1-1-2016

Measuring the Quality of the Website User Experience

Jeff Sauro
University of Denver

Measuring the Quality of the Website User Experience

Abstract

Consumers spend an increasing amount of time and money online finding information, completing tasks, or making purchases. The quality of the website experience has become a key differentiator for organizations--affecting whether they purchase and their likelihood to return and recommend a website to friends. Two instruments were created to more effectively measure the quality of the website user experience to help improve the experience.

Three studies used Classical Test Theory (CTT) to create a new instrument to measure the quality of the website user experience from the website visitor's perspective. Data were collected over five years from more than 4,000 respondents reflecting on experiences with more than 100 websites. An eight-item questionnaire of website quality was created - the Standardized User Experience Percentile Rank Questionnaire (SUPR-Q). The SUPR-Q contains four factors: usability, trust, appearance, and loyalty. The factor structure was replicated across three studies, with data collected both during usability tests and retrospectively in surveys. There was evidence of convergent validity with existing questionnaires, including the System Usability Scale (SUS). An initial distribution of scores across the websites generated a database used to produce percentile ranks and make scores more meaningful to researchers and practitioners. In Study 4, a new set of data and confirmatory factor analysis (CFA) confirmed the factor structure and generated alternative items that work on non-e-commerce websites. The SUPR-Q can be used to generate reliable scores in benchmarking websites, and the normed scores can be used to understand how well a website scores relative to others in the database.

A fifth study was designed to develop and evaluate guidelines regarding the quality of the user experience that could be judged by experts. Study 5 establishes a Calibrated Evaluator's Guide (CEG) for evaluators to review websites against a set of guidelines to predict perceptions of quality of website user experience. The CEG was refined from 105 to 37 items using the many-faceted Rasch model. The CEG was found to complement the SUPR-Q by providing a more detailed description of the website user experience. Suggestions for practical use and future research are discussed.

Document Type

Dissertation

Degree Name

Ph.D.

Department

Quantitative Research Methods

First Advisor

Kathy E. Green, Ph.D.

Second Advisor

Donald Bacon

Third Advisor

Duan Zhang

Keywords
Measurement, Usability, User experience

Subject Categories
Library and Information Science

Measuring the Quality of the Website User Experience

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements

for the Degree

Doctor of Philosophy

---

by

Jeff Sauro

June 2016

Advisor: Dr. Kathy Green

Author: Jeff Sauro
Title: Measuring the Quality of the Website User Experience
Advisor: Dr. Kathy Green
Degree Date: June 2016

ABSTRACT

Consumers spend an increasing amount of time and money online finding information, completing tasks, or making purchases. The quality of the website experience has become a key differentiator for organizations—affecting whether they purchase and their likelihood to return and recommend a website to friends. Two instruments were created to more effectively measure the quality of the website user experience to help improve the experience.

Three studies used Classical Test Theory (CTT) to create a new instrument to measure the quality of the website user experience from the website visitor's perspective. Data were collected over five years from more than 4,000 respondents reflecting on experiences with more than 100 websites. An eight-item questionnaire of website quality was created—the Standardized User Experience Percentile Rank Questionnaire (SUPR-Q). The SUPR-Q contains four factors: usability, trust, appearance, and loyalty. The factor structure was replicated across three studies, with data collected both during usability tests and retrospectively in surveys. There was evidence of convergent validity with existing questionnaires, including the System Usability Scale (SUS). An initial distribution of scores across the websites generated a database used to produce percentile ranks and make scores more meaningful to researchers and practitioners. In Study 4, a new set of data and confirmatory factor analysis (CFA) confirmed the factor structure and generated alternative items that work on non–e-commerce websites. The SUPR-Q can be used to generate reliable scores in benchmarking websites, and the normed scores can be used to understand how well a website scores relative to others in the database.

A fifth study was designed to develop and evaluate guidelines regarding the quality of the user experience that could be judged by experts. Study 5 establishes a Calibrated Evaluator’s Guide (CEG) for evaluators to review websites against a set of guidelines to predict perceptions of quality of website user experience. The CEG was refined from 105 to 37 items using the many-faceted Rasch model. The CEG was found to complement the SUPR-Q by providing a more detailed description of the website user experience. Suggestions for practical use and future research are discussed.

ACKNOWLEDGEMENTS

I want to first thank my advisor, Dr. Kathy Green for her diligent comments, patience and unparalleled responsiveness in answering questions and providing guidance. I’d also like to thank my committee: Dr. Duan Zhang, Dr. Krystyna Matusiak and Dr. Donald Bacon, for their time in providing thoughtful comments and flexibility during the review process. I’d like to thank the staff at MeasuringU for donating the data for this analysis. Finally I’d like to thank my wife, Shannon, for her patience and support through this process.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION AND LITERATURE REVIEW 1

Problem statement *4*
Research questions *5*
Literature review *6*
Usability testing *7*
Surveys *8*
Inspection methods *9*
Constructs and instruments to measure the quality of the user experience *10*
Synthesis of existing constructs and questionnaires *19*
Effectiveness of inspection methods *21*
Definitions *23*

CHAPTER 2 METHOD 26

Introduction *26*
Study 1: Item creation *26*
Study 2: Item refinement *28*
Study 3: Further item refinement and normalization *31*
Study 4: Confirmatory factor analysis *32*
Study 5: Calibrated Evaluator's Guide to User Experience Quality (CEG) *36*

CHAPTER 3: RESULTS 40

Study 1 .... *40*
Study 2 .... *45*
Study 3 .... *48*
Study 4 .... *55*
Study 5 .... *62*

CHAPTER 4: DISCUSSION 83

REFERENCES 95

APPENDIX A: Nielsen's 10 Heuristics 108

APPENDIX B: Tasks Used in Study 4 111

APPENDIX C: Tasks Used in Study 5 112

APPENDIX D: CEG Checklist 115

APPENDIX E : Comparison of CEG to Usability.gov Guidelines 121

APPENDIX F: Websites and Number of Evaluators Using CEG 125

APPENDIX G: Facets Output 128

APPENDIX H: Websites Used in the Rasch Analysis 137

LIST OF TABLES

Table 1. Questionnaires That Measure Aspects of Software and Website Quality, Total Number of Items, and Reported Reliabilities by Overall and Subscale Constructs 12

Table 2. Item Loadings and Communalities for the 13 Items..42

Table 3. Cronbach's Alpha and Minimum Inter-Item Correlations for the Four-Factor Solution 43

Table 4. Correlations between Factors, the Overall Score, and the System Usability Scale Score 44

Table 5. Factor Loadings for the Rotated Factor Solution for the 13 Candidate Items.....45

Table 6. Correlations between Subscales and the SUS and WAMMI Questionnaires......46

Table 7. Internal-Consistency Reliability Estimates (Cronbach's Alpha) and Minimum Inter-Item Correlations for the Four Factor Solution...47

Table 8. Factor Loadings for the Rotated Factor Solution for the 11 Candidate Items.....49

Table 9. Factor Loadings for the Rotated Factor Solution for the Eight Remaining Items 50

Table 10. Correlations between Subscales, Overall Score and the SUS Done at the Individual Response Level…………………………………………………………………..51

Table 11. Correlations between Subscales, Overall Score and the SUS Done at the Study Level (Averaged across Respondent by Website) 52

Table 12. Internal-Consistency Reliability Estimates (Cronbach's Alpha) and Minimum Inter-Item Correlations for the Four Factor Solution from the Eight Remaining Items…………………………………………………………………………………...53

Table 13. Mean and Standard Deviations for 2,513 Responses across 70 Websites by Overall Score and Subscale Scores...54

Table 14. Fit Indices for Four Tested Models..55

Table 15. Comparisons of Fit Indices for the Four Models...59

Table 16. Correlations between SUPR-Q, SUPR-Q Subscales, and Aws 60

Table 17. SUPR-Q Average Responses by Item and Experts (E1 – E3) Ratings.............61

Table 18. Evaluator Scores from the CEG, SUPR-Q Scores and Subfactors for 16 Websites...63

Table 19. Correlations between CEG Scores and SUPR-Q Raw Scores, Percentile Scores and Factors 64

Table 20. Retained 18 Items Relative to the 48 Used Initially in Winsteps 68

Table 21. Logit Positions by Item Retained..82

Table G-1. Judge Measurement Report (Arranged by MN) 110

Table G-2. Site Measurement Report (Arranged by MN)……………………………...115

LIST OF FIGURES

Figure 1. Distribution of subscales and overall score for 2,513 responses across 70 websites...54

Figure 2. Path diagram for Model 1..56

Figure 3. Path diagram for Model 2..57

Figure 4. Path diagram for Model 3..58

Figure 5. Path diagram for Model 4..59

Figure 6. Rating scale use..66

Figure 7. Category probability curves for responses to the CEG. 67

Figure 8. Dimensionality indices...68

Figure 9. Winsteps person fit...70

Figure 10. Winsteps dimensionality output. 71

Figure 11. Scale use for items and persons retained..71

Figure 12. Scale usage for retained items 72

Figure 13. Item map...73

Figure 14. Probability curves for many-faceted CEG model. 75

Figure 15. Item measurement report (arranged by MN)......................................76

Figure 16. Probability curves for collapsed three-point scale...77

Figure 17. All Facet vertical “rulers”...78

Figure 18. Item distribution for the 37 items in the CEG 79

Figure 19. Selection of output for Judge Measurement Report 81

Figure 20. Judge’s Experience Measurement Report (arranged by MN) 81

Figure G-1. Judge’s Experience Measurement Report (arranged by MN)……………......114

Figure G-2. Item Measurement Report (arranged by MN)……………………………..118

CHAPTER 1
INTRODUCTION AND LITERATURE REVIEW

The amount of money consumers spend online is substantial and increasing. In 2015, approximately $80 billion was spent online in the United States alone, representing 7 percent of all commerce in the United States (United States Census Bureau, 2015). US online retail sales are expected to grow at an annual rate of 9.5 percent through 2018 (Enright, 2014). Online consumers have many choices when making purchases or finding information on websites. If a user cannot find information, purchase a product easily, or does not trust the information, the user goes elsewhere and may tell friends and colleagues about the poor experience. Even for noncommercial websites like those for state and local governments, difficulty finding information or difficulty completing tasks often means increased cost from calls to call centers or the inability of citizens to fully utilize services. The quality of the website experience has, therefore, become a key differentiator for organizations. For example, while Walmart, the largest US retailer, currently derives only 3 percent of its $330 billion in revenue from online sales, it is relying heavily on its e-commerce business to drive growth, since its in-store sales have stagnated (Trefis Team, 2014). Amazon.com recently became one of the top US retailers, with over $44 billion in online sales, despite not having any physical stores—showing the importance of the online user experience (Stone, 2015). The quality of the user experience also affects whether people purchase. A survey of online customers revealed that poor website navigation was the reason 25 percent of respondents did not purchase

(Statistica, 2012). Corroborating these findings, earlier research found that navigation features that reduce the time to purchase products online account for 61 percent of the variance in online monthly sales (Lohse & Spiller, 1999). Attitude toward a website has been found to be a significant predictor of whether customers will revisit (Flores, 2004; Hong & Kim, 2004; Supphellen & Nysveen, 2001) and purchase again (Abdul-Muhmin, 2010). Repeat customers are more likely to purchase than first time visitors, and repeat visitors spend almost twice as much as first time visitors (George, 2002; Korgaonkar & Wolin, 2002).

The most common ways of evaluating the quality of the website experience are through collecting users' attitudes about the experience, observing users attempt tasks in a controlled setting on the websites (called a usability test), and having interface experts evaluate a website using guidelines and heuristics (Lewis, 2012b). A common method for evaluating users' attitudes is through the use of questionnaires administered after task completion or at the conclusion of a usability study (Sauro & Lewis, 2009) or through a survey administered on a website or emailed directly to recent website visitors. Standardized usability questionnaires, as opposed to homegrown questionnaires, have been shown to provide a more reliable measure of a user's experience (Hornbæk, 2006). However, standardized questionnaires alone are not particularly effective at diagnosing problems because they do not provide behavioral data that can illustrate the problems users encounter while using an interface (Sauro & Lewis, 2012). The types of questions asked are usually at too general a level to isolate particular issues (e.g., "The website is easy to use"). However, they are one of the most efficient ways of gauging the perception of an experience using measures that can most easily be compared across disparate products and domains.

The purpose of this study is to report the results of the development of two complementary instruments that measure the quality of the experience website visitors may have while browsing or purchasing on a website. The construct to be measured is termed “the quality of the website user experience.” The first instrument assesses several critical aspects of users’ judgment of a website experience. Users are asked to respond to items such as “I think this website is easy to use” or “I am able to find what I need quickly” after having some experience using the website. It will be called the SUPR-Q (Standardized User Experience Percentile Rank Questionnaire). For this measure to be useful, it needs to be short to minimize the burden on participant users; contain a reference database to bring more meaning to the scores; and include questions specific to the website user experience but not so specific that they are irrelevant on disparate types of websites (e.g., nonprofit versus e-commerce websites).

This second instrument is designed to be used by trained evaluators instead of users. It contains more items to provide more concrete details on the strengths and weaknesses of websites—as they influence user behavior. For example, evaluators will judge how well a website conforms to statements such as “Button and link labels effectively indicate where the user will be taken.” Such items are derived from common problems users encounter while using websites and that often inhibit purchases or locating information. It will be called the Calibrated Evaluator’s Guide to User Experience Quality (CEG).

Because it can be difficult to collect data from website users (especially on less popular websites), this longer instrument can be used by evaluators who judge how well a website contributes to or hinders the actual user experience. The evaluators assess common areas of websites known to affect user attitudes, such as the navigation, product pages, labels, search functionality, and checkout pages. This separate instrument provides a complementary view of the quality of the website user experience to the one collected directly from users. Thus, the intent of this study was to develop a measure of quality of the user experience that can be used by visitors to a website (the SUPR-Q) and also a quality of user experience measure that can be used by website evaluators (the CEG).

Problem statement

A measure of the quality of the user experience on websites helps diagnose interaction problems and generates a benchmark against which to test design improvements or feature enhancements. The new instrument needs to be:

- Generally useful: It needs to provide enough dimensions to sufficiently describe the quality of a website experience but not be so specific that it cannot be used with many different types of websites. For example, information websites differ from e-commerce websites, which in turn differ from nonprofit websites. Item phrasing needs to be sufficiently generic that the same items can be used.
- Multifaceted: It needs to encompass the most well-defined factors for measuring website quality as uncovered in the review of existing instruments and review of the literature.

- Brief: It needs to be brief, since time with participants is precious, and with the increase in mobile usage, answering lengthy questionnaires on small screens is prohibitive.
- Backed by a normative database: Finally, knowing where a website scores relative to its peers in a normative database will provide additional information to researchers who administer the instrument in isolation for a new website.

Although some existing instruments share some of these aspects, none contain all four aspects (i.e., short, covering multiple facets including trust, and with a normative database). The purpose of this study was to develop two instruments: one that measures the quality of a website user experience that is generalizable, multidimensional, brief, and backed by a normative database, and a second that was designed for use by independent judges to evaluate elements of a website based on guidelines that can provide more diagnostic and detailed information on what to improve.

Research questions

This research addresses the following questions, which pertain to one or both instruments:

1] What aspects best quantify the quality of the website user experience?

2] Which items have the best psychometric properties for measuring the construct of the quality of the website user experience?

3] Does the new instrument have sufficient reliability while still remaining short (for the SUPR-Q)?

4] Does the new instrument demonstrate adequate construct and content validity?

5] Can users' attitudes toward website quality as measured by a validated instrument be predicted from using a website user experience quality checklist (for the CEG)?

6] Does the experience level of the evaluator affect the ratings on the CEG?

Literature review

The terms "usability" and "user experience" are used somewhat interchangeably in practice but represent different but related constructs. There is disagreement about the actual difference and definitions (Stewart, 2015). Usability refers to the ability of participants to complete tasks effectively and efficiently and is embodied in an international standard, ISO 9241 (part 11) (International Organization for Standardization [ISO], 1998). Nielsen offers a similar definition with usability comprising five quality components: learnability, memorability, efficiency, error prevention, and satisfaction (Nielsen, 1993).

In contrast, user experience is a broader term that includes usability but also the more ethereal constructs of beauty, hedonic, affective, or experiential aspects of a technology (Hassenzahl & Tractinsky, 2006). It can be applied to products where one wants to have a good experience but effectiveness and efficiency are not a primary concern, such as with video games. Under this distinction, a product or website can be usable but offer a poor user experience—usability is necessary although not sufficient for a good user experience.

Despite their differences, both usability and the user experience are primarily measured using similar evaluation methods (Tullis & Albert, 2013). The website user experience is most commonly measured using one of three methods: observing participants attempt tasks in a controlled setting (called a usability test), collecting users' attitudes about a prior experience in a survey, or having interface experts evaluate a website using guidelines and heuristics—often called inspection methods or expert reviews (Lewis, 2012b).

Usability testing

In a usability test, a representative set of participants is recruited and asked to attempt realistic tasks in a controlled environment, often in a dedicated "lab" co-located with a facilitator and often a note-taker and other observers (Dumas & Redish, 1999). The main outcome of a usability test is a test report. Usability test reports contain lists of usability problems and metrics that describe the quality of the experience. Usability problems are the most common output of a usability test and form the basis of recommendations on what to improve (Nielsen, 1993; Rubin & Chisnell, 1994). Usability problems describe the interaction of elements in an interface (e.g., labels, organization, messaging) that lead to errors, longer task times, and failed task attempts, and consequently less usage or purchasing (Lohse & Spiller, 1999; Nielsen, 1993). Preventing common usability problems is the intent of guidelines that experts use to inspect an interface (covered in detail in the latter part of this section).

Usability test reports also contain performance metrics (e.g., task completion rates, task times, errors) and questionnaires about users' attitudes toward the task and overall product or website experience (Sauro & Lewis, 2009). Many of the user experience questionnaires contain standardized measures of the overall experience that provide a valuable benchmark to gauge the effectiveness of future design changes and can be compared across disparate products and websites (Whiteside, Bennett, & Holtzblatt, 1988).

Usability testing, however, is costly and time consuming. While expensive facilities and equipment are not necessary to conduct a test, it still requires the time of a skilled facilitator and time and resources for participants to be recruited, compensated, and usually co-located with facilitators. Consequently, usability test sample sizes are often small, and tests are reserved for when critical problems need to be diagnosed and corrected (Krug, 2014; Sauro, 2010b). Test metrics and user attitudes have a lower priority than finding and fixing usability problems in usability tests (Sauro & Lewis, 2009).

Surveys

A more cost effective way than usability testing to collect attitudinal metrics is through surveying website users and asking them to reflect on their recent experience. Surveys can be sent to many users via email or as a website "pop-up" intercept that can be used to collect data from a large sample quickly. While these surveys often contain similar questions about the overall quality of the experience as those used in usability tests, they do not provide more detailed performance metrics (task completion, errors, and times) or usability problems because there is no observed behavior. Despite not having detailed behavioral data to help guide website improvements, these self-reported measures do provide a valuable gauge of the website user experience and provide a high-level idea about what areas are problematic. Surveys are the more popular method for wide-scale measurement of the quality of the website user experience, since there are low costs and quick turnaround to electronically collecting data from customers.

Inspection methods

While surveying users is usually more cost effective than conducting usability tests, both methods still present challenges. Even with short instruments, collecting data from website users can be time consuming and often expensive. For many websites with a low number of daily visitors, it can take months to collect a sufficient number of responses using website intercept surveys like those deployed by ForeSee.com.

There is a rich history in the usability literature on using analytic methods, as opposed to empirical methods, to uncover problems in an interface (Hollingsed & Novick, 2007). Popular methods include some variety of an expert reviewing the interface: heuristic evaluations (Nielsen & Molich, 1990), cognitive walkthroughs (Lewis, Polson, Wharton, & Rieman, 1990), and guideline reviews (Bastien & Scapin, 1997).

In a heuristic evaluation, an expert in usability principles reviews an interface against a set of broad principles called heuristics. For example, one of the most common sets of heuristics is the 10 from Nielsen (1994). The heuristics were derived from an examination of many problems uncovered in usability tests to generate overall principles. For example, one heuristic is “The system should speak the users’ language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms” (Nielsen, 1994, p. 156). All 10 heuristics are shown in Appendix A (Nielsen, 1995). The expert then inspects the website to determine how well it conforms to this heuristic and identifies shortcomings (Nielsen, 1993).

A cognitive walkthrough is a usability inspection method like heuristic evaluation, but the emphasis is put on task scenarios users would likely perform with the software or website (Lewis et al., 1990). Users' goals and how they would attempt to accomplish the goals in the interface are first identified. Then, an expert in usability principles identifies problems users might encounter as they learn to use an interface by meticulously going through each step.

A guideline review involves having an evaluator compare an interface against a detailed set of guidelines. Guidelines can be used both for creating an interface (typically used by designers and developers) or evaluating it for compliance (typically performed by usability evaluators). Guideline reviews predate the web and became more popular with the increase in graphical user interfaces (GUIs). One of the best-known and most comprehensive was a set of guidelines sponsored by the US Air Force and MITRE Corporation. Published in 1986, *Guidelines for Designing User Interface Software* by Smith and Mosier contains 944 mostly usability-related guidelines (Smith & Mosier, 1986). Later, Apple released their *Human Interface Guidelines* (Apple Computer, 1987) and Microsoft followed (Microsoft Corporation, 1995).

In the subsequent sections, I review the extant literature on measures of user experience quality and then guidelines and techniques that predict problems in interfaces without requiring data collection from users.

Constructs and instruments to measure the quality of the user experience

The concept of user experience quality predates the web. Standardized usability questionnaires first appeared in the late 1980s and are widely used today (Sauro & Lewis, 2012). Those first questionnaires were technology-agnostic, meaning the items were appropriate for software, hardware, or any physical device. The advantage of a technology-agnostic instrument is that the scores obtained can be compared regardless of the technology used to gather them. An organization can use the same set of scores to benchmark mobile applications as well as desktop interfaces. The disadvantage of a technology-agnostic instrument is that it can omit important information that is specific to an interface type.

Measurement of user experience quality gained wider use with the proliferation of desktop software, with some of the first instruments operationalizing quality in terms of technology acceptance (feature utility) and ease of use (Davis, 1989). A number of complementary instruments measuring technology adoption, including the Technology Acceptance Model (TAM), were found to be strong predictors of future use (Venkatesh, Morris, Davis, & Davis, 2003). For a more detailed history of software quality measurements, see Sauro and Lewis (2012).

There are a number of published instruments that measure various aspects of website quality. Details about them, including subscales, number of items, and reliabilities, are listed in Table 1 and are briefly described below. The most commonly used instruments are technology-agnostic and were developed before the web as we know it existed (e.g., Chin, Diehl, & Norman, 1988; Davis, 1989).

Table 1

Questionnaires That Measure Aspects of Software and Website Quality, Total Number of Items, and Reported Reliabilities by Overall and Subscale Constructs

Questionnaire	# Items	Measures	Global Reliabilit y	Sub-Constructs	Construct Reliabilit y	Source
SUS	10	System Usability	0.92	Usability	0.91	Brooke (1996)
				Learnability	0.71	Borsci, Federici, & Lauriola (2009); Sauro & Lewis (2009)
PSSUQ	16	Perceived Satisfaction	0.94	System Quality	0.9	Lewis (1992)
				Information Quality	0.91	
				Interface Quality	0.83	
SUMI	50	Usability	0.92	Efficiency	0.81	Kirakowski (1996)
				Affect	0.85	
				Helpfulness	0.83	
				Control	0.71	
				Learnability	0.82	
QUIS	27	Interaction Satisfaction	0.94	Overall Reaction	nr	Chin et al. (1988)
				Screen Factors	nr	
				Terminology and System Feedback	nr	
				Learning Factors	nr	
				System Capabilities	nr	

WAMMI	20	Website Usability	0.90	Attractiveness	0.64	Kirawoski & Cierlik (1998)
				Controllability	0.69	
				Efficiency	0.63	
				Helpfulness	0.70	
				Learnability	0.74	
WQ	25	Website Quality	0.92	Specific Content	0.94	Aladwani and Palvia (2002)
				Content Quality	0.88	
				Appearance	0.88	
				Technical Adequacy	0.92	
WU	8	Website Usability	nr	Ease of Navigation	0.85	Wang and Senecal (2007)
				Speed	0.91	
				Interactivity	0.77	
IS	15	Information Satisfaction	nr	Customer Centeredness	0.92	Lascu and Clow (2008)
				Transaction Reliability	0.80	
				Problem-Solving Ability	0.77	
				Ease of Navigation	0.61	
ISQ	13	Intranet Satisfaction	0.89	Content Quality	0.89	Bargas-Avila, Lötscher, Orsini, & Opwis (2009)
				Intranet	0.90	

				Usability		
UMUX	4	Perceived Usability	0.94	Perceived Usability	0.94	Finstad (2010)
UMUX-LITE	2	Perceived Usability	0.82	Perceived Usability	0.82	Lewis, Utesch, & Maher (2013)
HQ	7	Hedonic Quality	nr	Ergonomic Quality	nr	Hassenzahl (2001)
				Appeal	nr	
ACSI	14-20	Customer Satisfaction	nr	Quality	nr	www.theasci.org
				Freshness of Information	nr	
				Clarity of Site Organization	nr	
				Overall Satisfaction	nr	
				Loyalty	nr	
CxPI	3	Customer Experience	nr	Usefulness	n/a	www.forrester.com
				Usability	n/a	
				Enjoyability	n/a	
NPS	1	Customer Loyalty	n/a	Customer Loyalty	n/a	Reichheld (2003)
TAM	12	Technology Acceptance	nr	Usefulness	0.98	Davis (1989)
				Ease of Use	0.94	
WebQual	36	Website Quality	nr	Informational Fit to Task	0.86	Loiacono, Watson, & Goodhue (2002)
				Tailored Communication	0.80	
				Trust	0.90	
				Response	0.88	

Time	
Ease Of Understanding	0.83
Intuitive Operations	0.79
Visual Appeal	0.93
Innovativeness	0.87
Emotional Appeal	0.81
Consistent Image	0.87
On-Line Completeness	0.72
Relative Advantage	0.81

Table Notes: nr = Not Reported, n/a = not applicable. Reliability values are Cronbach’s alpha.

The 10-item System Usability Scale (SUS), developed by Brooke (1996), is perhaps the most frequently used questionnaire to measure perceived usability across products and websites (Sauro & Lewis, 2009). While the SUS was not published with a normative database, enough data have been collected and enough of it published that it is possible to create a set of normed scores (Sauro, 2011). Tullis and Stetson (2004) found the SUS to be the best discriminating questionnaire of websites’ usability. A more recent scale for measuring usability is the Usability Metric for User Experience (UMUX) developed by Finstad (2010). At just four items, it is reliable and short. Lewis et al. (2013) used a two-item variation, called the UMUX-LITE, which was also found to be reliable and correlated highly with the SUS.

Other frequently used technology-agnostic instruments for measuring perceived usability include the Post Study System Usability Questionnaires (PSSUQ) (Lewis, 1992), the Software Usability Measurement Inventory (SUMI) (Kirakowski, 1996), and the Questionnaire for User Interaction Satisfaction (QUIS) (Chin et al., 1988). The SUMI contains a reference database maintained by its authors, but at 50 items, the instrument is the longest among those identified.

There are other instruments that measure factors other than usability. A standardized questionnaire to measure website quality and related constructs is the Website Analysis and Measurement Inventory (WAMMI) (Kirawoski & Cierlik, 1998). The current version of the WAMMI has a set of 20 items covering the five subscales of attractiveness, controllability, efficiency, helpfulness, and learnability and the global WAMMI measure. The WAMMI, like the SUMI, contains a reference database based on data collected from users of the questionnaire and maintained by its authors. Users of the WAMMI can convert their raw score into a percentile rank based on scores from the other websites in the database. The internal consistency reliability of the WAMMI global score is high (α =.90), whereas the subscale reliability estimates are generally lower (α =.63 to α =.74). The lower reliability is a tradeoff for using fewer items to measure a construct (Bobko, 2001). The WAMMI uses four items to measure each of five constructs. Brevity is often critical when participants' time is limited, so the loss in reliability arguably can be justified by higher response rates and adoption of the instrument. Information about the number and type of websites in the database is not provided in the authors' reports, but this slightly shorter multifactor instrument with a reference database is a model for the current research. The database behind the WAMMI makes it appealing to generate comparison scores as can be done with the Customer Experience Index (CxPI) developed by the consulting firm Forrester. The CxPI consists of only three items measuring usefulness, usability, and enjoyability. There is, however, no published information on the psychometrics of the CxPI (www.forrester.com).

Wolfinbarger and Gilly (2002) developed a four-factor model of website quality specific to the online purchasing experience that correlated with loyalty and customer satisfaction. Their 14-item questionnaire, called .comQ, includes factors of website design, reliability, privacy/security, and customer service. A longer 40-item version provides more concrete areas to fix and addresses one of the common shortcomings of shorter instruments—not having enough diagnostic information for website developers. The .comQ builds on earlier qualitative work by Zeithaml, Parasuraman, and Malhotra (2000), who identified 11 dimensions of the online experience: access, ease of navigation, efficiency, flexibility, reliability, personalization, security/privacy, responsiveness, assurance/trust, site aesthetics, and price knowledge.

Lo Storto (2013) created a model of website efficiency as a measure of quality and performance. Fifty-two e-commerce websites were evaluated using a three-factor, nine-item questionnaire that includes the dimensions of user experience, site navigability, and structure. He found that inefficient websites over-utilize specific inputs (e.g., force the users to make a greater cognitive effort during navigation) or under-produce outputs (e.g., provide the users with scarce gratification when they use the websites).

While websites may be treated under the broader category of software, they bring the very salient elements of trust and visual appeal into consideration. Bevan (2009) argues that to encompass the overall user experience, measures of website satisfaction need to account for likability and trust. Other researchers have also found that online trust is a major determinant of e-commerce success (Keeney, 1999; Pavlou & Fygensen, 2006; Suh & Han, 2003).

Safar and Turner (2005) developed a psychometrically validated trust scale consisting of two factors based on an online insurance quote system. A broader examination of website trust was also conducted by Angriawan and Thakur (2008). They found that website usability, expected product performance, security, and privacy collectively explained 70 percent of the variance in online trust. They also found that online trust and privacy were strong predictors of consumer loyalty, which was similar to findings by Sauro (2010a) and Lewis (2012a).

The WebQual questionnaire by Loiacono et al. (2002) is a more comprehensive (but longer) 36-item measure that contains subscales including trust, usability, and visual appeal. The construct of visual appeal appears in multiple questionnaires, including the WAMMI. The instrument by Aladwani and Palvia (2002) contains an appearance subscale, and the influential Hedonic Quality (HQ) questionnaire developed by Hassenzahl et al. (2001) also has an appeal subscale. Additional instruments focus on narrower aspects of website quality, specifically satisfaction, including questionnaires by Wang and Senecal (2007), Lascu and Clow (2008), and Bargas-Avila et al. (2009).

Customer loyalty plays an important role in business decisions and appears as a construct in multiple questionnaires. The most popular is the Net Promoter Score (NPS). The NPS is a single 11-point scale (0 to 10) intended to measure customer loyalty (Reichheld, 2003). Respondents are asked to rate how likely they are to recommend a friend or colleague to a product or service. Responses of 0 to 6 are considered “detractors”; 7 to 8, “passives”; and 9 to10, “promoters.” The proportion of detractors is subtracted from the proportion of promoters to create the “net” promoter score. Research conducted by Reichheld (2006) showed that the NPS was the best or second best predictor of company growth in 11 out of 14 industries (not just limited to web-based industries). It is used widely across many industries, and benchmark data are available from third-party providers. Its high adoption rate makes it a good candidate for inclusion in this instrument. Similar loyalty measures appear in website questionnaires from the American Customer Satisfaction Index (ACSI) maintained by the University of Michigan (www.theasci.org) and by the company ForeSee, a proprietary instrument with no published reliabilities or details but which is used by many websites (ForeSee.com).

Synthesis of existing constructs and questionnaires

In reviewing the literature, the most common constructs related to measuring the quality of the user experience are usability (including navigation ease), trust, appearance, and loyalty. Some research (e.g., Sauro [2010a], and Lewis [2012]) suggests that these are overlapping constructs, since they were found to be correlated (e.g., trust and usability and usability and loyalty). These constructs will form the basis of the definition of quality of the user experience in the current study and were considered in developing items used on a new website questionnaire.

There are four reasons existing measures cannot be used and a new measure is needed. Some are proprietary instruments (e.g., ForeSee, WAMMI); many do not provide enough coverage of website quality (e.g., only usability with the SUS or loyalty with the NPS); others are too long (e.g., WebQual, WAMMI, SUMI, SUS); and many have insufficiently documented psychometric properties (e.g., CxPI).

Website-specific guidelines review

Some organizations have developed website-specific guidelines. For example, Nielsen has 113 guidelines for homepage usability (Nielsen, 2001), and Stanford has 10 guidelines for credibility (Stanford University, 2004). A formal effort at creating website guidelines was made by the US Department of Health and Human Services (HHS). HHS created a set of 209 guidelines specifically for websites (United States Department of Health and Human Services [HHS], 2006). Each guideline is based on published research, with many guidelines providing both examples and peer-reviewed references to support the guidelines. The guidelines were further evaluated by 13 experts (PhDs with background in usability and experimental design) and rated on importance using a five-point importance scale with the anchors of 1 = "Important" to 5 = "Very Important." They reported internal consistency reliability of the importance ratings (Cronbach's alpha of .92).

A more recent, albeit less formal, set of website guidelines was put together by David Travis of Userfocus.uk, a UK-based usability consulting firm (Travis, 2009). The 247 guidelines are available in a downloadable spreadsheet and have become popular with usability evaluators. The guidelines were created at a more granular level than the HHS guidelines. Travis explains:

> The spreadsheet came about as a way for me to structure my own expert reviews. Although usability guidelines like Nielsen's and ISO's are valuable, I find that students find them too high level to make yes/no decisions. So I created these design guidelines based on various sets of principles (Nielsen, ISO,

> Shneiderman), issues I saw as causing problems in usability testing and a general view of “good practice.” Some are much more important than others so I encourage people not to think of it as gospel, but more as an *aide memoire* to remind them what to look for while reviewing (D. Travis, personal communication, August 1, 2015).

Effectiveness of inspection methods

There are a number of studies comparing the effectiveness of each inspection method and the number of problems uncovered vis-à-vis usability tests (e.g., Jeffries, Miller, Wharton, & Uyeda, 1991; John & Marks, 1997; Karat, Campbell, & Fiegel, 1992). One theme that emerged in these studies was the high variability in results between evaluators. Nielsen and Molich (1990) warned that any single evaluator is unlikely to uncover most of the usability problems. They recommended using between three and five evaluators. More recent research has found that multiple evaluators conducting heuristic evaluations independently tend to find between 30 and 50 percent of the problems also found in a concurrently run usability test (Law & Hvannberg, 2004; Sauro, 2012).

There is evidence that using more detailed guidelines improves the quality of inspection methods. Bastien and Scapin (1995) found that evaluators following guidelines uncovered more problems that those who just inspected the interface. They argued that ergonomic-based guidelines can act as a framework for evaluators by reducing the variability and increasing the ability to detect issues. Jeffries et al. (1991)

found that a guidelines-based approach forces a more careful examination of the interface relative to heuristic evaluations or cognitive walkthroughs.

An early instrument developed by Chen and Wells (1999), based on the advertising literature, was created for evaluators instead of users. At least three judges (in this case judges were drawn from 120 MBA and undergraduate students) evaluated 120 websites based on six items (e.g., comfort, website relationship building, intentions to revisit, satisfaction with service). The instrument, called the A_{st} (Attitude toward the Site) was shown to be unidimensional. The students also rated 65 adjectives that correlated with the A_{st}. Three factors (entertainment, informativeness, and organization) accounted for the majority of variance in the A_{st} scores. In a follow-up study, Chen, Clifford, and Wells (2002) used website developers instead of students as judges on a new set of websites. They replicated the unidimensionality of the A_{st} and tested a new trust dimension. They found that the same three factors best explained A_{st} scores and that trust did not contribute additional explanatory value. They concluded that students were sufficient surrogates for more experienced (and expensive) judges of website quality.

Bruner and Kumar (2002) created a three-item Attitude toward the Website scale (Aws) and found that it discriminated better than did the Chen and Wells (1999) A_{st}. The three items are “I think it’s a good website,” “I liked the website,” and “I think it’s a nice website.” This finding was echoed by Boostrom, Balasubramanian, and Summey (2013).

Summary

This chapter reviewed the literature pertaining to the concept of the quality of the website user experience, which found the common attributes of website quality being usability, trust, appearance, and loyalty. In reviewing the instruments, their subfactors, and their corresponding reliabilities, a corpus of items was identified in the literature for the creation of the new quality of user experience instrument. A review of the literature on guidelines revealed that guideline reviews, when used in conjunction with inspection techniques, can uncover problems in the quality of an interface. There was no research showing that such guidelines could predict users' attitudes, however.

The intent of this study was to develop two measures of the quality of the website user experience. The first measure was administered to website users, was brief, and was backed by a normative database. The second instrument was used to evaluate elements of a website based on guidelines evaluated by independent judges that provides more diagnostic and detailed information on what to improve. Existing measures are either not comprehensive enough, do not have a normative database (or are proprietary), are too long, or do not address diagnostic needs.

The design for creating the new instrument of website quality and guidelines to predict users' attitudes toward website quality are discussed in Chapter Two. Chapters Three and Four contain a summary of the results and the general conclusions derived from the study along with recommendations for future research.

Definitions

The following is a list of common terms used throughout this dissertation.

Usability: The "ease of use" of an interface. It is defined formally in the ISO 9241 definition (part 11) (ISO, 1998) as the extent to which an interface can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.

User Experience: Overlapping but broader than usability, the user experience encompasses more than just the ability of an interface to allow users to accomplish tasks quickly and efficiently. It is also defined in the ISO 9241 definition (part 210) as the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors, and accomplishments that occur before, during, and after use.

Usability Test: An empirical evaluation method that involves watching users attempt tasks on an interface (website, app, or product) to identify problems (formative) or measure its ease of use with metrics (summative).

Inspection Method: A family of analytic usability evaluation techniques (as opposed to empirical ones like usability testing and surveys) where a set of experts evaluate an interface, often against criteria, that identify potential problems for users.

Heuristic Evaluation: A type of inspection method where evaluators judge an interface against a set of general design principles, or heuristics, to identify potential problems for users. See Appendix A for examples of 10 heuristics.

Guideline Review: A type of inspection method where evaluators examine an interface to see how well it conforms to detailed guidelines that should make an experience more usable.

SUS: The System Usability Scale (SUS) is the most commonly used instrument for measuring perceptions of a product or website’s usability. It is a 10-item instrument typically administered after a usability test.

Quality of Website User Experience: The experience and attitudes a user has while attempting to accomplish tasks to achieve specific goals on a website.

CHAPTER 2
METHOD

Introduction

In this chapter, four studies are described that led to creation of the new instrument to measure the quality of the website user experience from the website visitor's perspective (SUPR-Q). In Study 1, an initial pool of items informed by those described in the literature review was identified and tested with an initial convenience sample. In Study 2, the items were refined using a larger sample size and more websites to establish reliability and convergent validity with existing instruments was estimated. In Study 3, the basis of the normalized database was created using a larger sample size and a reduced number of items and continued evidence of validity and reliability was compiled. In Study 4, a new set of data and confirmatory factor analysis (CFA) were used to confirm the factor structure and explore alternative items. A fifth study was designed to develop and evaluate guidelines regarding the quality of the user experience that could be judged by experts. Study 5 establishes a questionnaire (CEG) for evaluators to review websites against a set of guidelines to predict perceptions of quality of website user experience.

Study 1: Item creation

Purpose. The purpose of Study 1 was to create an initial pool of items for the development of the instrument and to assess initial reliability and validity.

Participants. Initial data were collected via a convenience sample of adults who shop online and have made a purchase in the last six months. The initial sample size goal was 100 surveys to have a reasonable number of responses to support performing a factor analysis. One hundred surveys were completed that contained responses from around the United States with a mix of gender (60 percent female) and average age of 34 (27 to 63).

Instrument. An initial set of 33 items was constructed corresponding to the four constructs of usability, loyalty, trust, and appearance (based on their ability to describe website quality). A five-point response scale (strongly disagree = 1 to strongly agree = 5) was used, except for the item "How likely are you to recommend the website to a friend," which used a 0-to-10-point scale. By keeping a 0 to 10 scale, this item can be used to compute the Net Promoter Score (Reichheld, 2003).

The 10-item System Usability Scale (SUS) with a five-point response scale (strongly disagree = 1 to strongly agree = 5) was used (Brooke, 1996) for establishing convergent validity. The SUS has high internal consistency reliability (alpha = .92) and has been shown to differentiate between usable and unusable software applications (Sauro, 2011).

Procedure. An email was sent to friends and colleagues of the author (US-based participants). Participants were asked to reflect on their most recent online purchasing experience and answer an initial pool of candidate items plus the SUS (Brooke, 1996) items in an online survey. Respondents were asked from which website they completed their purchase and what they purchased.

Analysis. Classical test theory (CTT) was used to assess the psychometric properties of the pilot version of the new instrument (Nunnally, 1978).

An exploratory factor analysis (EFA) using principal factor analysis without rotation was first conducted to determine if the data were factorable; parallel analysis was used to determine how many factors to retain. Based on the overlapping constructs in the literature, it was anticipated that factors would be correlated. If the data were considered factorable, an EFA would be conducted using an oblique rotation (direct oblimin) and only the highest-loading items retained. While the cutoff for retaining items is based somewhat on the preference of the researcher, Tabachnick and Fidell (2001) recommend a minimum .32 loading. To further winnow the number of items down, items with low item–total correlations were removed. To allow for reliability analysis of the subscales, it is necessary to retain a minimum of two items per factor. Thus, the shortest possible questionnaire that assesses four factors would have to be eight items.

Reliability evidence for the scale was assessed using internal consistency reliability with Cronbach’s alpha. To assess convergent validity of the usability subfactor, System Usability Scale (SUS) total score was correlated with a composite average score of the retained candidate items.

Study 2: Item refinement

Purpose. A second study was conducted with the candidate items retained in Study 1 to replicate the factor structure, to continue to reduce the number of items, and to initiate a normalized dataset. Study 2 used a larger sample size per website, plus a broader range of websites.

Participants. A larger and more diverse sample of US-based participants was surveyed. The participants were adults (18+ years of age) who browse and purchase products online, primarily from the United States. A total of 484 surveys were completed. Between 10 and 15 users attempted a task on one of the websites with a mix of gender (58 percent female), median age of 33 (18 to 68), a mix of occupations (including professionals, homemakers, and students), mix of education levels (45 percent bachelors, 34 percent high school/GED, and 18 percent advanced degree) representing 47 states. Participants had a range of experience with each website, with the lesser-known poor-quality websites having no users who had prior experience, compared with moderate exposure for some participants on the higher-traffic websites.

Instrument. To further assess the convergent validity of the instrument, the WAMMI questionnaire was used in eight websites' surveys, and the SUS was also included along with the candidate items from Study 1.

The SUS was described under Study 1 (above). The 20-item Website Analysis and Measurement Inventory (WAMMI) (Kirawoski & Cierlik, 1998) global score was used to assess convergent validity. The WAMMI consists of five subscales, attractiveness, controllability, efficiency, helpfulness, and learnability, as well as the global WAMMI measure. The internal consistency reliability of the WAMMI global score is high (α =.90) (Kirawoski & Cierlik, 1998).

Procedure. Participants were recruited using online advertisements, Amazon's Mechanical Turk, and panel agencies to participate in a short online survey. Participants were compensated to complete the survey with a target sample size of 400 to 500 (at least 10 responses per website). A screening question was included in the survey (i.e., Asking participants to select response option 3) to filter out participants who might be rushing through the survey to collect the honorarium. Participants were asked to attempt one predefined task on one of 40 websites and answer the candidate items selected from Study 1 (and the WAMMI and SUS for a subset of the websites). The websites tested represented a range of quality, with some having known poor-quality experiences and others having known high-quality experiences. The poor-quality websites were selected from the website webpagesthatsuck.com. The high-quality websites were from some of the most visited websites in the United States. They came from a range of industries, including retail, travel, information technology, government, and cellular service carriers. Examples of poor-quality websites include: 1001pens.com, NY State Government, Tally-Ho Uniforms, Julie Garwood, and Crumpler. Higher-quality websites include: Expedia, Sprint, Target, Walmart, and Budget. The tasks participants attempted were tailored for each website (e.g., finding airline ticket prices, locations of government offices, or product prices).

Analysis. An exploratory factor analysis (EFA) using principal axis factoring without rotation was first conducted to determine if the data were factorable, and parallel analysis was used to decide how many factors to retain. It was anticipated that there would be four factors and that the factors would be correlated. If the data were considered factorable, an EFA would be conducted using an oblique rotation (direct oblimin) and only the highest-loading items retained. To further winnow the number of items down, items with low item–total correlations were removed.

Finally, to assess the ability of the instrument to discriminate between poor and good experiences, an ANOVA was conducted using the average score of the items (dependent variable) with each website (independent variable).

Reliability evidence for the scale or subscales was assessed using Cronbach’s alpha. To assess convergent validity of the subscales, the SUS score and WAMMI score were correlated with subscale scores.

Study 3: Further item refinement and normalization

Purpose. Based on the results from Study 2, a third study was conducted with a larger sample size to continue to refine the factors and begin to establish a normalized database of website scores.

Participants. To qualify, participants must have visited or made a purchase on one of 70 websites in the prior six months. The participants were from the general US internet population, with a mix of genders and who actively browse or purchase products online. The targeted sample size was between 30 and 100 responses per website, for a total minimum sample size of 2,100. A total of 3,891 surveys were completed, with a mix of gender (53 percent female), median age of 29 (18 to 73), a mix of occupations (including professionals, homemakers, and students), mix of education levels (46 percent bachelors, 40 percent high school/GED, and 5 percent advanced degree). Participants had a range of experience with each website, with each participant having had to visit the website at least once in the prior six months. Participants reflected on their experience with 51 websites across a range of industries, including Delta, Craigslist, USA.gov, eBay, The New York Times, Office Depot, and Walgreens. The median number of responses per website was 44 (14 to 126 responses).

Instrument. The remaining candidate items from Study 2 were used along with items asking for core demographic information and prior experience with the websites being used. The SUS was administered and was described on page 31.

Procedure. Participants were recruited using online advertisements, Mechanical Turk, and panel agencies to participate in a short online survey. Participants were compensated to participate in the survey. A screening question was included in the survey (i.e., Asking participants to select response option 3) to filter out participants who might be rushing through the survey to collect the honorarium. Unlike Study 2, where participants were randomly assigned to different websites and asked to complete a task, Study 3 asked participants to reflect on their recent experience with one of 70 websites. Participants were asked to respond to one website, even if they visited multiple websites. Additionally, data from cognitive interviews and comments from participants using the instrument were used to see if additional final changes in item wording were indicated.

Analysis. Composite subscale scores from the instrument were again correlated with the SUS to assess convergent validity. Reliability evidence for the scale was assessed using Cronbach's alpha. Websites that received "at least usable" responses were compiled into a normalized database.

Study 4: Confirmatory factor analysis

Purpose. The purpose of Study 4 was to first confirm the factor structure of the newly created questionnaire using a new dataset and to test alternative items on the trust factor. The second purpose was to assess the convergent validity of the new items with another validated instrument and then examine content validity with the final items by correlating the average scores by item with independent expert ratings.

Participants. To qualify, participants must have made purchases online at least monthly. Participants must be 18+ years of age and browse or purchase products online. For the CFA sample, a total of 2,093 surveys were completed with a mix of gender (55 percent female), median age of 32 (18 to 54) and mix of household income (58 percent between $25,000 and $75,000 USD).

For the convergent validity sample, a total of 151 surveys were completed with a mix of gender (61 percent male), median age of 33 (18 to 59), from across the United States. For the content validity sample, three experts in website and user experience measurement were asked to judge how difficult each item was for respondents to agree to. Each of the experts had more than 20 years of experience in the field of usability and interface evaluation. Each had been published multiple times, and one expert had published multiple psychometrically validated instruments on measuring the user experience. The three experts were male, two from the United States, one from the United Kingdom, and all three were older than 55 years of age.

Instrument. Participants were asked to answer the items on the newly developed instrument along with additional candidate items and standard demographic information such as gender and age. Ten items were included. Seven of the items were taken from Study 3, two alternative items were added which were phrased more generally about trust (ideal for websites without a purchasing component), and one item that had similar psychometric properties from Study 2 but that did not include a purchasing component was included. The 10 items are listed below along with the facet the item addresses:

1. The website is easy to use. (usability)

2. It is easy to navigate within the website. (usability)
3. The website keeps the promises it makes on its website (trust from Study 2)
4. I feel confident conducting business on the website. (trust)
5. The information on the website is credible. (candidate trust)
6. The information on the website is trustworthy. (candidate trust)
7. How likely are you to recommend this website to a friend or colleague? (loyalty)
8. I will likely return to the website in the future. (loyalty)
9. I find the website to be attractive. (appearance)
10. The website has a clean and simple presentation. (appearance)

The three items from the Attitude toward the Website scale (Aws) were used to establish convergent validity (Bruner & Kumar, 2002). The three items are "I think it's a good website," "I liked the website," and "I think it's a nice website."

For establishing content validity, the experts were asked to rate how easily participants would find it to agree to each of the final items in the SUPR-Q using a three-point scale (1 = Easy to Agree To; 2 = Medium to Agree To; 3 = Hard to Agree To).

Procedure. For the CFA, participants were recruited using a panel agency to participate in an online survey about their attitude toward popular e-commerce websites. Participants were paid between $3 and $10 to complete the survey. A survey link was sent to the panel participants and the data were collected in an online survey tool. Participants were assigned to one of five retail websites (e.g., eBay, Walmart), asked to look for items of interest, and then asked to answer the candidate items.

For the convergent validity sample, participants were recruited using Amazon's Mechanical Turk and assigned to one of five real estate websites (Realtor, Zillow, Trulia, Redfin, or Homefinder) to complete a task or to view one of four video streaming websites (Vudu, Amazon Instant Video, HBO Go, YouTube). A screening question was again included in the survey (i.e., Asking participants to select response option 3) to filter out participants who might be rushing through the survey to collect the honorarium. The tasks are listed in Appendix B. Participants were paid between $3 and $10 to complete the survey. For the content validity sample, the three experts were asked to consider each item independently (each expert lives in different cities, two in the United States and one in the United Kingdom) and submit their responses in an Excel worksheet.

Analysis. A confirmatory factor analysis (CFA) was used to assess both a one-level factor structure and a hierarchical structure. The higher-order factor used an overall quality of the user experience factor with subfactors derived from analyses in Studies 1, 2, and 3.

Using CFA, the fit of the model was assessed using a chi-square test. The chi-square test is known to be overly sensitive to sample size. Thus, the Comparative Fit Index (CF), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR) fit indices were used to further assess model fit (Kline, 2011). For non-nested models, the Akaike’s Information Criterion (AIC) and its family of fit indices were used to differentiate fit adequacy between the more complex and simpler models (Kline, 2011).

For assessing convergent and content validity, the Pearson and Spearman correlations between scores were used to estimate validity.

Study 5: Calibrated Evaluator’s Guide to User Experience Quality (CEG)

Purpose. The purpose of this study was to identify a pool of items from existing guideline sources to create a quality of user experience measure that can be used by website evaluators. This Calibrated Evaluator’s Guide (CEG) pulls from research-based guidelines available at HHS (HHS, 2006) and userfocus.uk (Travis, 2009) to create a measure that is more diagnostic than the SUPR-Q.

Participants. Two types of participants were enlisted. The first group was two evaluators who independently rated between eight and 16 websites using the newly created guideline measure. Both evaluators were new to usability evaluation but had several weeks training in evaluating interfaces and reviewing guidelines and checklists.

The second group involved participants recruited using Amazon’s Mechanical Turk and who were compensated to participate in the study. The participants were US-based adults who browse or purchase products online. A total of 322 completed questionnaires were returned from 68 websites representing data from 225 different evaluators.

Instrument. The guidelines created by HHS, while comprehensive, contain many guidelines that may be difficult to evaluate effectively and are similar to the ones in Travis (2009). For example, Guideline 1.1, “Provide Useful Content,” is ranked highest in importance and in strength of evidence. The full guideline is “Provide content that is engaging, relevant, and appropriate to the audience.”

Redundant and potentially confusing guidelines (e.g., “ The site avoids marketing waffle”) were removed or rewritten from the initial list, which created a new set of 105 guidelines (see Appendix D). Appendix E shows the overlap between the new guidelines and the HHS guidelines. The major difference was that the new set contained specific guidelines toward product pages and purchasing, whereas the HHS guidelines were broader and did not focus on these often critical aspects of the e-commerce website experience. The new set of guidelines was grouped in the eight categories listed below and are shown in Appendix D:

1. Navigation
2. Information
3. Search
4. Product Pages
5. Purchasing and Billing
6. Forms & Data Entry
7. Help & Information
8. Overall Elements

Evaluators were instructed to rate how well the website conformed to the guideline using a five-point scale from Strongly Disagree (1) to Strongly Agree (5). The average of the 105 items becomes an overall score for a website.

Procedure. To assess convergent validity, participants were recruited using Mechanical Turk and randomly assigned to one of 16 websites (listed in Appendix C) and asked to either complete the CEG guidelines or attempt two tasks on the website (See Appendix C for the task list). A screening question again was included in the to filter out participants who might be rushing through the survey to collect the honorarium After the tasks, the participants responded to the SUPR-Q or the items in the CEG. The targeted sample size was 25 to 27 usable responses per website, or 400 total responses. This provided 16 SUPR-Q scores and CEG total scores to assess a correlation between users' website user experience and evaluators' judgment of the website user experience.

An advantage of using a few trained evaluators to complete a guideline-based questionnaire is that they ostensibly have a better ability to accurately rate the experience and yield more reliable ratings. A disadvantage is that the judgment of a few evaluators can be skewed by prior experience with brand or website. To minimize the biases introduced by using a few of the same evaluators, between five and 10 participants recruited from Mechanical Turk were assigned to one website and asked to answer a portion of the guideline-based questionnaire. Participants were paid between $1 and $2 to answer between 12 and 70 of the 105 items in the questionnaire. These results were then compared against the group of more experienced evaluators.

Analysis. To estimate convergent validity, the guideline-based questionnaire was correlated with the score on the SUPR-Q developed and refined in Studies 1 through 4. The correlations were based on independent data sources—different participants used the CEG and the SUPR-Q. Interrater reliability was also computed between two independent evaluators' CEG scores on the 16 websites (Appendix C) using the Pearson correlation.

The Rasch model (Rasch, 1980) was used for additional scale analysis to go beyond rating consistency to identify how the type of evaluator interacted with the items. A diverse set of websites was evaluated that offer a range of content and user experience quality. Multiple evaluators were assigned a subset of the CEG on one of 52 new websites selected from the ForeSee annual website Customer Satisfaction benchmark (ForeSee, 2015). They include government websites and retail websites.

To understand the interaction between evaluators and items, a many-faceted Rasch analysis was conducted on the responses to identify the items that best fit the model. In a many-faceted Rasch analysis, each ordinal observation is conceptualized to be the outcome of an interaction between the items and evaluators (Linacre, 1989). That is, each evaluator is expected to be more lenient or severe when judging the website against the items in the CEG. The Rasch model is extended to incorporate the severity or leniency of the judge. The model was assessed for unidimensionality by examining principal components of the residuals (Linacre, 1989). The items were assessed for fit to the model using infit (weighted) and outfit (unweighted) mean square fit statistics. Poorly fitting judges were removed if infit mean squares exceeded 3.0 and items were removed if infit mean squares exceeded 1.5.

CHAPTER 3: RESULTS

Study 1

A total of 100 surveys were completed, but nine surveys contained at least one missing value, leaving 91 fully completed surveys. In total, 51 unique websites were listed, with the most responses coming from Amazon (33), eBay (five), and Bn.com (four).

An exploratory factor analysis using principal factor analysis without rotation was conducted to determine if the data were factorable and if so, how many factors to retain. The Kaiser–Meyer–Olkin Measure of Sampling Adequacy was .86 and Bartlett's Test of Sphericity was statistically significant, χ^2 (528) = 2191.37, *p* <.001, supporting factorability (Tabachnick & Fidell, 2001). A scree plot of the eigenvalues suggested a three-, four-, or five-factor solution.

A parallel analysis was also conducted, with data showing three factors with eigenvalues greater than those from randomly simulated matrices. While the parallel analysis suggested retaining only three factors, this initial sample size was small, relative to the items being considered, and there was a theoretical rationale to look at four correlated constructs of website quality (usability, trust, appearance, and loyalty).

Given that factors were likely to be correlated, an exploratory factor analysis using principal axis factoring with oblique rotation (direct oblimin) was then conducted and four factors were retained. Items with factor loadings less than .5 were removed (this removed seven items). The four factors were named loyalty, trust & credibility, usability, and appearance based on the item content for items that loaded on each factor.

The remaining 26 items were broken out into their corresponding factors, and a reliability analysis was conducted for each factor. In keeping with the goal of a parsimonious instrument, items were winnowed down to as few as possible per factor. For each factor, items with item–total correlations less than .5 and with cross-loadings on multiple factors within .2 were deleted. Of the remaining items, those with the highest factor loadings and highest item–total correlation were retained, leaving three to four items per factor. A few items had negatively worded tones, and those were dropped to keep an all-positive instrument to avoid coding and interpretation problems (Sauro & Lewis, 2011).

The exploratory factor analysis was rerun using principal axis factoring with oblimin rotation to extract four factors. The factors, items, and communality are shown in Table 2. A total of 13 items remained.

Table 2

Item Loadings and Communalities for the 13 Items

	Usability	Trust	Loyalty	Appearance	Communality
I am able to find what I need quickly on this website.	**0.88**	0.10	-0.02	0.15	0.81
It is easy to navigate within the	**0.87**	-0.09	0.02	-0.05	0.78
This website is easy to use.	**0.80**	-0.13	0.09	-0.01	0.67
I feel comfortable purchasing from this website.	0.02	**-0.92**	-0.06	0.00	0.84
This website keeps the promises it makes to me.	-0.05	**-0.89**	0.05	0.04	0.80
I feel confident conducting business with this website.	0.10	**-0.89**	0.06	-0.11	0.82
I can count on the information I get on this website.	0.02	**-0.88**	-0.07	0.07	0.78
I consider myself a loyal customer of this website.	0.01	-0.03	**0.94**	-0.05	0.89
How likely are you to recommend this website to a colleague or friend?	-0.09	-0.02	**0.87**	0.12	0.78
I plan on continuing to purchase from this website in the future.	0.16	0.07	**0.81**	-0.02	0.69
I found the website to be	0.03	0.04	-0.09	**0.93**	0.87
The website has a clean and simple presentation.	0.15	-0.04	0.04	**0.78**	0.63
I enjoy using the website.	-0.06	-0.14	0.35	**0.67**	0.59
Eigenvalue	5.92	2.35	1.33	0.95	
% of Variance	45.51	18.06	10.25	7.29	
Cumulative %	45.51	63.57	73.82	81.11	

Note that to allow for reliability analysis of the subscales, it is necessary to retain a minimum of two items per factor. Thus, the shortest possible questionnaire that assesses four factors would have eight items. The internal consistency reliability estimates and minimum inter-item correlations are shown in Table 3. All subscales showed reliabilities above .70 (Nunnally, 1978).

Table 3

Cronbach's Alpha and Minimum Inter-Item Correlations for the Four-Factor Solution

	Cronbach's Alpha	Minimum Inter-Item Correlation
Appearance	.83	.60
Loyalty	.83	.58
Usability	.87	.67
Trust	.93	.70
Overall	.87	.12

To assess the convergent validity of the candidate subscales, scores on each subscale were averaged and correlated with the 10 SUS items, along with a composite score created by averaging all 13 items. The correlations are shown in Table 4.

Table 4

Correlations between Factors, the Overall Score, and the System Usability Scale Score

	Usability	Trust	Loyalty	Appearance	Overall
Trust	0.68				
Loyalty	0.36	0.32			
Appearance	0.54	0.38	0.50		
Overall	0.77	0.77	0.73	0.74	
SUS	0.59	0.36	0.64	0.73	0.71

All correlations were statistically significantly different from zero at the $p < .01$ level. The usability, loyalty, and appearance factors all correlated at between $r = .59$ and .73 with the System Usability Scale. The overall composite score correlated at $r = .71$ with SUS. These medium-to-high correlations suggest convergent validity with the SUS. The medium-to-high correlations between factor average scores also confirm the correlation between factors as suggested in the literature and support the use of an oblique rather than an orthogonal rotation. The different correlations between the factors and SUS are expected given that SUS was meant to measure only usability. However, a higher correlation between the SUS and the appearance factor suggests that attitudes toward website appearance and usability are comingled. For further discussion on the relationship between website usability and appearance, see Tuch, Roth, Hornbæk, Opwis, and Bargas-Avila (2012).

Study 2

A total of 484 surveys were completed. To assess the factor structure, principal axis factoring using oblimin rotation was conducted. The Kaiser–Meyer–Olkin Measure of Sampling Adequacy was .92 and Bartlett’s Test of Sphericity was statistically significant, χ^2 (78) = 4015.20; $p < .001$. The factor loadings and communalities are shown in Table 5.

Table 5

Factor Loadings for the Rotated Factor Solution for the 13 Candidate Items

	Trust	Usability	Appearance	Loyalty	Communality
I can count on the information I get on this website	**0.95**	0.03	0.05	-0.12	0.91
I feel confident conducting business with this website.	**0.73**	-0.18	-0.01	0.08	0.56
This website keeps the promises it makes to me.	**0.73**	-0.04	0.04	-0.01	0.53
I feel comfortable purchasing from this website.	**0.59**	-0.23	-0.09	0.20	0.45
The information on this website is valuable.	**0.48**	0.10	0.06	0.17	0.27
I am able to find what I need quickly	0.00	**-0.87**	0.00	0.07	0.76
This website is easy to use.	0.09	**-0.84**	0.01	0.06	0.71
It is easy to navigate within the website.	0.05	**-0.74**	0.24	-0.01	0.61
I found the website to be attractive.	0.08	0.06	**0.71**	0.14	0.54
The website has a clean and simple presentation.	0.04	-0.29	**0.67**	-0.03	0.53
I will likely purchase something from this website in the future.	0.01	-0.01	0.06	**0.69**	0.48

How likely are you to recommend the website to a friend or colleague?	0.14	-0.19	0.07	**0.57**	0.39
I enjoy using the website.	0.09	-0.23	0.25	**0.40**	0.28
Extraction Sums of Squared Loadings	7.45	0.88	0.49	0.32	
% of Variance	57.33	6.79	3.78	2.46	
Cumulative %	57.33	64.11	67.89	70.36	
Rotation Sum of Squared	5.92	5.52	4.90	4.90	

In examining the factor loadings in Table 5, the items still fit a four-factor structure reasonably well with most loadings above .6. To further reduce the number of items, two items, “I enjoy using the website” and “The information on this website is valuable,” that had the lowest loading on their respective factors were dropped.

To assess the convergent validity of the four subscales, scores were created by averaging the item scores for each subscale and correlating them with SUS (n = 441) and WAMMI (n = 106). The correlations are shown in Table 6.

Table 6

Correlations between Subscales and the SUS and WAMMI Questionnaires

	Usability	Trust	Loyalty	Appearance	Overall	SUS
Trust	0.66					
Loyalty	0.64	0.67				
Appearance	0.68	0.64	0.63			
Overall	0.87	0.87	0.88	0.82		
SUS	0.88	0.71	0.69	0.73	0.87	
WAMMI	0.86	0.71	0.66	0.67	0.85	0.95

All correlations were statistically significant at $p < .01$. The usability score and overall score showed the highest convergent validity with strong correlations with both the SUS ($r \geq .87$) and WAMMI ($r \geq .85$). All subscales were, however, significantly and moderately to strongly correlated with both the SUS and WAMMI.

The internal consistency reliability estimates for each subscale and minimum inter-item correlations are shown in Table 7. All subscales and the overall scale showed reliabilities to be above .70, except for the loyalty factor with a coefficient alpha of .63.

For some websites used in the study, participants could not make a purchase, rendering the item “I will likely purchase something from this website in the future” irrelevant. A more generic version of this item, “I will likely visit this website again in the future,” will be used in subsequent analysis and may increase the internal consistency reliability. The measure created is called the SUPR-Q (Standardized User Experience Percentile Rank Questionnaire).

Table 7

Internal-Consistency Reliability Estimates (Cronbach’s Alpha) and Minimum Inter-Item Correlations for the Four Factor Solution

	Cronbach’s Alpha	Minimum Inter-Item Correlation
Appearance	.82	.69
Loyalty	.63	.61
Usability	.94	.85
Trust	.89	.44
Overall	.91	.39

For eight websites, the SUS, WAMMI, and SUPR-Q were collected for 108 total responses. A one-way analysis of variance was used with the SUPR-Q total score as the dependent variable and website as the independent variable with eight levels. A significant effect was found for website, *F*(7,100) = 4.43, *p* < .001 (Adj r-square = 18.34%). The SUPR-Q exhibited the same or equal differentiating power in terms of score by website as the SUS, *F*(7,100) = 4.52, *p* < .001 (Eta-squared = .24) and WAMMI, *F*(7,100) = 4.22, *p* < .001 (Eta squared = .23). The average of the three items on the usability factor differed by website, *F*(7,100) = 5.19, *p* < .001 (Eta-squared = .27), as did the loyalty subscale, *F*(7,100) = 5.57, *p* < .001 (Eta-squared = .28); to a lesser extent, the trust score differed by website, *F*(7,100) = 2.91, *p* =.008 (Eta-squared = .17); and appearance did not significantly differ, F(7,100) = 1.39, *p* = .22.

The 13 items loaded as expected on a four-factor structure. Two of the items had low loadings and were dropped. The remaining 11 items showed high overall internal consistency reliability and sensitivity in discriminating between websites with poor and high quality. The reliability of the subscales was high, with the exception of the loyalty factor, which had a coefficient alpha below .70, below the generally acceptable cutoff (Nunnally, 1978). Finally, a few participants in the survey comments noted that the item “This website keeps the promises it makes to me” sounded awkward. One participant commented for example, “How can a website keep promises?” Rewording the item to “The website keeps the promises it makes” was tested in the subsequent study.

Study 3

A total of 3,891 surveys were completed. To assess the factor structure with this new set of data, a principal axis factoring using oblique rotation was conducted. The Kaiser–Meyer–Olkin Measure of Sampling Adequacy was .93 and the Bartlett’s Test of

Sphericity was statistically significant, χ^2 (55) = 26897.4, p <.001. The factor loading matrix is shown in Table 8.

Table 8

Factor Loadings for the Rotated Factor Solution for the 11 Candidate Items

	Usability	Trust	Loyalty	Appearance
It is easy to navigate within the website.	**0.88**	0.01	0.00	0.00
The website is easy to use.	**0.87**	0.01	0.02	-0.01
I am able to find what I need quickly on the website.	**0.58**	0.09	0.12	0.11
I feel comfortable purchasing from the website.	0.02	**0.86**	-0.07	0.01
I feel confident conducting business on the website.	0.06	**0.84**	0.05	-0.04
I can count on the information I get on the website.	-0.01	0.35	0.31	0.20
I will likely return to the website in the	0.05	-0.01	**0.78**	-0.03
How likely are you to recommend the website to a friend or colleague?	0.04	-0.02	**0.77**	0.04
The website keeps the promises it makes to	-0.01	0.35	0.39	0.16
I find the website to be attractive.	-0.01	0.01	0.03	**0.76**
The website has a clean and simple presentation.	0.34	-0.01	-0.03	**0.56**
Extraction Sums of Squared Loadings	5.90	0.91	0.44	0.20
% of Variance	53.67	8.25	3.97	1.80
Cumulative %	53.67	61.92	65.88	67.68
Rotation Sums of Squared Loadings	4.82	3.89	4.54	4.62

Three items were dropped. The item “I am able to find what I need quickly on the website” had the lowest relative loading on the usability factor and was dropped. The items “The website keeps the promises it makes to me” and “I can count on the information I get on the website" both had loadings below .4 and cross-loaded on multiple factors. This reduced the total number of items to eight.

To assess the factor structure with these eight items, principal axis factoring using oblique rotation was conducted. The Kaiser–Meyer–Olkin Measure of Sampling Adequacy was .86 and the Bartlett's Test of Sphericity was significant, χ^2 (28) = 17512, p <.001. The final factor matrix is shown in Table 9.

Table 9

Factor Loadings for the Rotated Factor Solution for the Eight Remaining Items

	Usability	Trust	Loyalty	Appearance
The website is easy to use.	**0.88**	0.02	0.02	-0.02
It is easy to navigate within the website.	**0.80**	0.02	0.03	0.06
I feel comfortable purchasing from the website.	-0.01	**0.87**	-0.05	0.02
I feel confident conducting business on the website.	0.03	**0.83**	0.08	-0.02
How likely are you to recommend the website to a friend or colleague?	-0.01	-0.01	**0.80**	0.05
I will likely return to the website in the future.	0.03	0.01	**0.79**	-0.03
I find the website to be attractive.	-0.05	0.03	0.05	**0.76**
The website has a clean and simple presentation.	0.25	0.00	-0.02	**0.64**
Extraction Sums of Squared Loadings	4.26	0.80	0.42	0.18
% of Variance	53.24	10.0	5.30	2.26
Cumulative %	53.24	63.3	68.60	70.85
Rotation Sum of Squared Loadings	3.53	2.77	3.26	3.47
Cronbach's Alpha for Scale	.88	.85	.64	.78

The final eight-item SUPR-Q reflects the multi-factor solution for measuring the quality of the user experience of websites and having a normalized database with more than 100 websites.

To assess the convergent validity of the eight-item SUPR-Q, scores were created by averaging the items for the global score and for each subscale for each participant's response (participant-level scoring). For a subset of the websites, responses to the 10-item SUS were also collected, and the global score and subscale scores were correlated at the participant level (n = 2513). The correlations are shown in Table 10 below.

Table 10

Correlations between Subscales, Overall Score and the SUS Done at the Individual Response Level

	SUS	SUPR-Q	Usability	Trust	Loyalty
SUPR-Q	0.75				
Usability	0.73	0.85			
Trust	0.39	0.62	0.46		
Loyalty	0.61	0.84	0.60	0.49	
Appearance	0.64	0.85	0.73	0.48	0.57

All correlations calculated at the participant level were statistically significantly different from zero ($p < .001$). The usability factor score and overall score showed high convergent validity with strong correlations with SUS ($r > .73$).

Correlations were calculated again at the study level by averaging the scores across participants (study-level coding) where the average score for each website was correlated (n = 40). It has been shown that study-level metrics tend to correlate higher than individual metrics (Sauro & Lewis, 2009), and it is the study-level scores that are of interest to researchers. The correlations are shown in Table 11 below.

Table 11

Correlations between Subscales, Overall Score and the SUS Done at the Study Level (Averaged across Respondent by Website)

	SUS	SUPR-Q	Usability	Trust	Loyalty
SUPR-Q	0.87				
Usability	0.87	0.88			
Trust	0.47	0.57	0.40		
Loyalty	0.82	0.91	0.73	0.72	
Appearance	0.73	0.86	0.81	0.71	0.64

All correlations at the study level were statistically significantly different from zero (p < .001). The usability factor score and overall score showed high convergent validity with strong correlations with SUS (r = .87). A reliability analysis was conducted on the four factors and the coefficient alpha and item correlations are shown in Table 12.

The overall composite score made up of eight items and the usability factors both had high reliability (coefficient alpha >.85), and the trust and appearance factors had acceptable reliability (coefficient alpha >.75), while the loyalty factor had low reliability (coefficient alpha = .64).

Table 12

Internal-Consistency Reliability Estimates (Cronbach's Alpha) and Minimum Inter-Item Correlations for the Four Factor Solution from the Eight Remaining Items

	Cronbach's Alpha	Minimum Inter-Item Correlation
Appearance	.78	.64
Loyalty	.64	.65
Usability	.88	.78
Trust	.85	.73
Overall	.86	.36

For 40 websites, the SUS and eight candidate items were collected for 2,513 total responses. A one-way ANOVA was used with the combined average score for the eight items as the dependent variable and website as the independent variable with 40 levels. The combined average score differed significantly between the poorest- and highest-quality websites, *F*(39,2473) = 10.22, *p* < .001 (Eta-squared = .14). It exhibited about equal differentiating power to the SUS, *F*(39,2473) = 9.67, *p* < .001 (Eta-squared = .13), with two fewer items. The subscales also provided evidence for sensitivity by differentiating between the websites on usability, *F*(39,2473) = 6.03, *p* < .001 (Eta-squared = .13); trust, *F*(39,2473) = 12.13, *p* < .001 (Eta-squared = .16); loyalty, *F*(39,2473) = 14.80, *p* < .001 (Eta-squared = .19); and appearance, *F*(39,2473) = 5.82, *p* < .001 (Eta-squared = .08).

The distribution of the average scores for the overall composite and the subscales for the 70 websites is shown in the histograms in Figure 1.

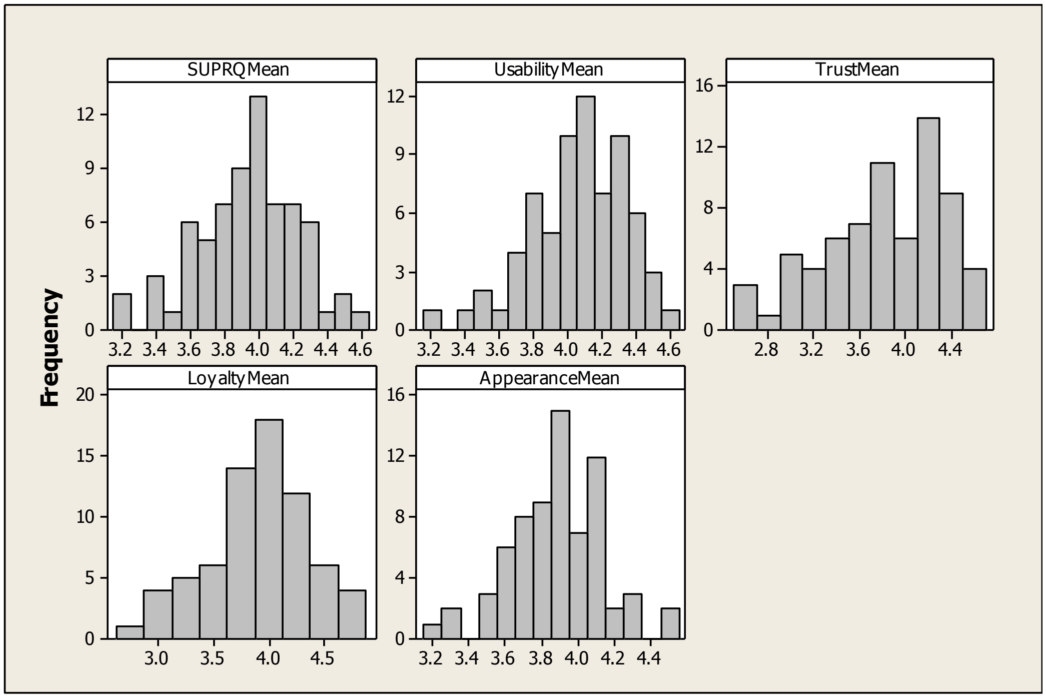


Figure 1. Distribution of subscales and overall score for 2,513 responses across 70 websites.

The distribution of scores is generally normal, with a skewness and kurtosis for the SUPR-Q mean of -0.30, 11; usability mean of -.56, .32; trust mean of -.50, -.50; loyalty mean of -.44, -.12; and appearance mean of -.10, .49. The means and standard deviations for each of the subscales and overall are shown in Table 13.

Table 13

Mean and Standard Deviations for 2,513 Responses across 70 Websites by Overall Score and Subscale Scores

	Mean	St**d.** Dev**.**	Skewness	Kurtosis
SUPR-Q	3.93	0.29	-0.30	0.11
Usability	4.06	0.29	-0.56	0.32
Trust	3.80	0.52	-0.50	-0.50
Loyalty	3.91	0.46	-0.44	-0.12
Appearance	3.88	0.25	-0.10	0.49

Study 4

A confirmatory factor analysis was conducted on four models assessing the fit of new items and a hierarchical versus flat structure. Table 14 shows the four models and their corresponding fit statistics. The intent of the analysis was to compare a higher order with a flat model with and without alternative items that did not reference purchasing or doing business on a website.

Table 14

Fit Indices for Four Tested Models

	Model 1	Model 2	Model 3	Model 4
Hierarchical	Higher Order	Flat	Flat	Higher Order
Items	Original Items	Original Items	Alt Items	Alt Items
Chi-Square	266.467	153.761	129.586	180.374
Sample Moments	36	36	36	36
Parameters	20	22	22	20
DF	16	14	14	16
SRMR	0.022	0.017	0.014	0.017
RMSEA	0.087	0.069	0.063	0.070
90% Low	0.078	0.059	0.053	0.061
90% High	0.096	0.079	0.073	0.079
RMSEA p	<.0001	0.001	0.015	<.0001
CFI	0.981	0.989	0.992	0.988
AIC	306.467	197.761	173.586	220.374
BIC	419.394	321.981	297.806	333.301

Model 1 assessed a higher order factor with the original eight items. The path diagram is shown in Figure 2.

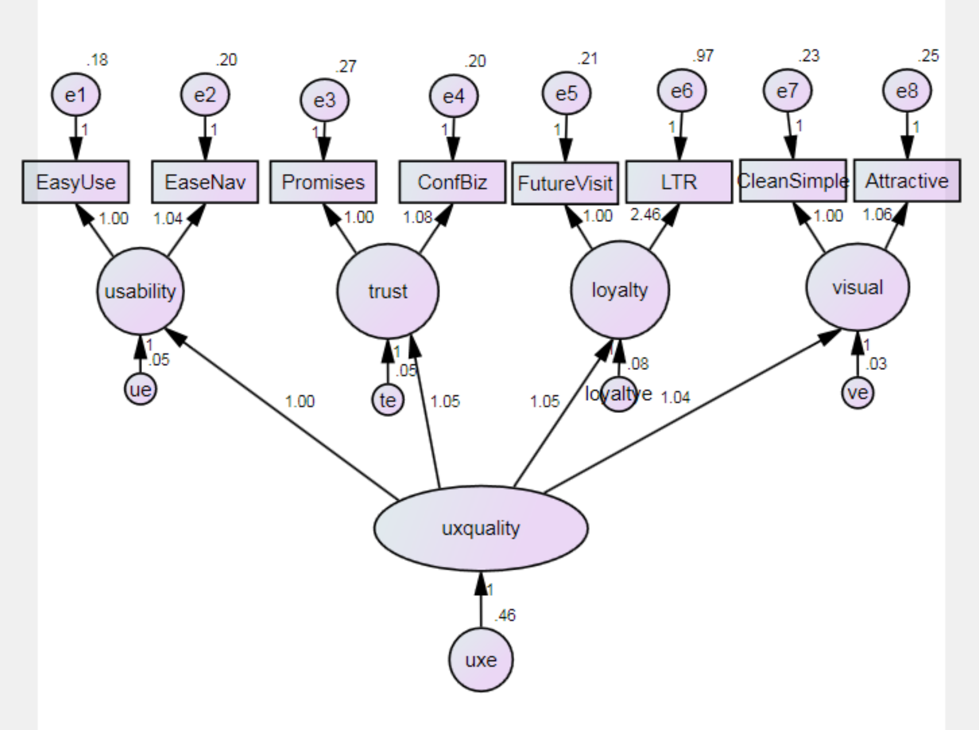


Figure 2. Path diagram for Model 1.

It is an over-identified model with 36 observations (sample moments) and 20 parameters, *df* = 16. The model chi-square test was statistically significant, $\chi^2 = 266.47$, $p < .01$, thus the null hypothesis of adequate fit was rejected. This suggests that the model did not fit well. However, chi-square is a conservative test, especially with large sample sizes, and often rejects the null hypothesis of perfect fit (Kline, 2011). In looking further at the model fit indexes, RMSEA was .081. The 90% CI was .078 to 0.087, suggesting good fit, as the upper bound is below .10. The SRMR value of .022 was less than .08, showing very good fit (Hu & Bentler, 1998). The CFI was .981, which is above .90 and also suggests good model fit (Kline, 2011).

Model 2 assessed the four factors from Study 3 (no higher-order factor) with the original eight items. The path diagram is shown in Figure 3.

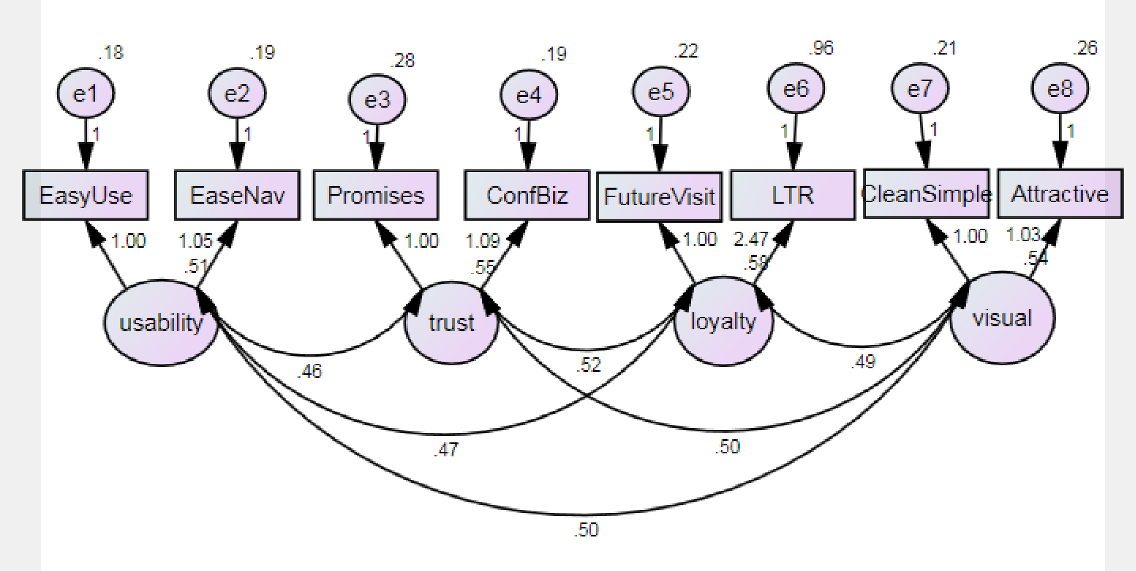


Figure 3. Path diagram for Model 2.

It is an over-identified model with 36 observations (sample moments) and 22 parameters, *df* =14. The model chi-square test was statistically significant, χ^2 = 153.761, *p* < .01. This suggests that the model does not fit well, which is again overly conservative. In looking further at the model fit indexes, RMSEA was .069 with a 90% CI of .059 to 0.079, suggesting good fit, as the upper bound is below .10. The SRMR value of .017 is less than .08, showing very good fit (Hu & Bentler, 1998). The CFI of .989 also suggests good model fit (Kline, 2011).

Model 3 assessed the four factors from Study 3 (no higher-order factor) with the alternate items, and the path diagram is shown in Figure 4.

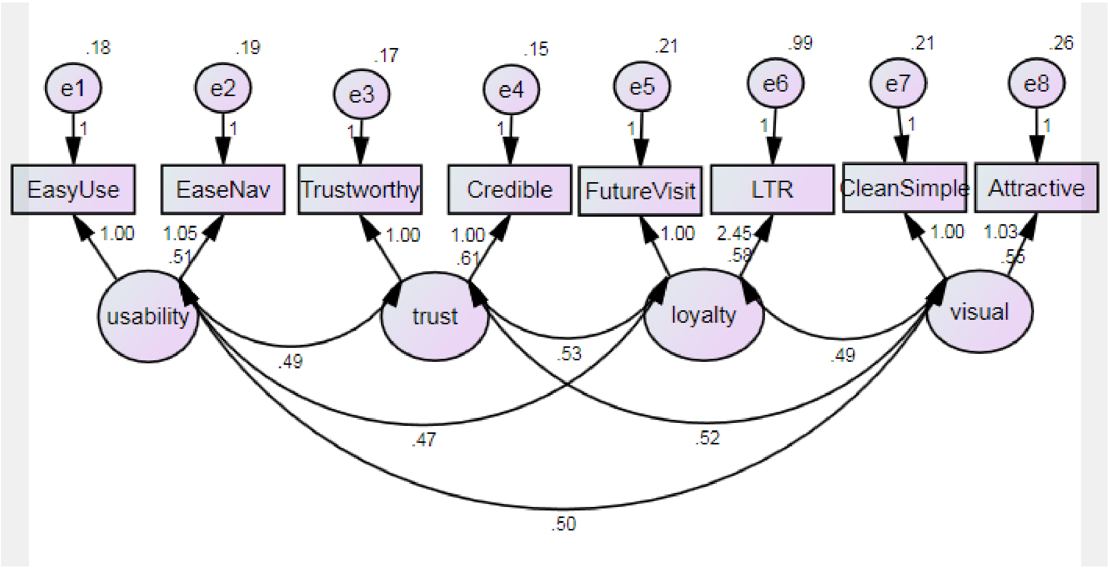


Figure 4. Path diagram for Model 3.

It is an over-identified model with 36 observations (sample moments) and 22 parameters, *df* = 14. The model chi-square test was statistically significant, χ^2 = 129.586, *p* < .01. In looking further at the model fit indexes, RMSEA was .063 with a 90% CI of .053 to 0.073, suggesting good fit as the upper bound is below .10. The SRMR value of .014 is less than .08, showing very good fit (Hu & Bentler, 1998). The CFI of .992 also suggests good model fit (Kline, 2011).

Model 4 assessed the four factors from Study 3 with a higher-order factor and the alternate items.

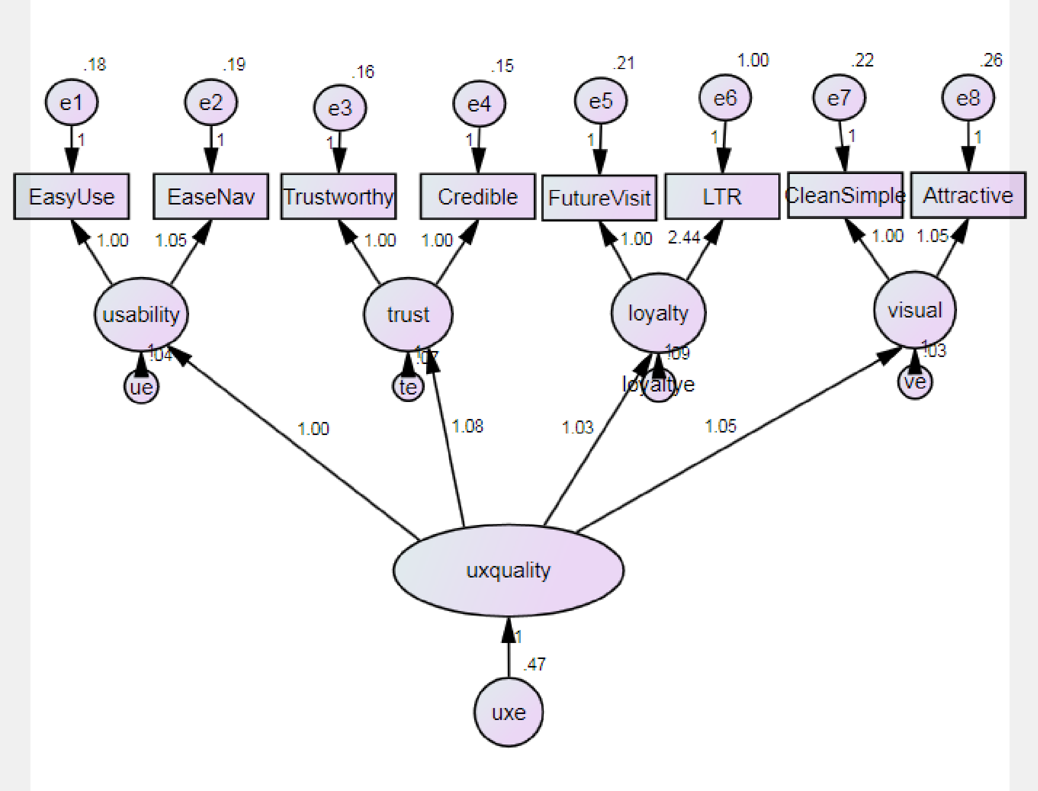


Figure 5. Path diagram for Model 4.

It is an over-identified model with 36 observations (sample moments) and 20 parameters, *df* = 16. The model chi-square test was statistically significant, χ^2 = 180.374, *p* < .01. RMSEA was .070 with a 90% CI of .061 to 0.079, suggesting good fit as the upper bound is below .10. The SRMR value of .017 was less than .08, showing very good fit (Hu & Bentler, 1998). The CFI of .988 also suggests good model fit (Kline, 2011).

Table 15

Comparisons of Fit Indices for the Four Models

Model Comparison	**Chi-Square Diff.**	**Df**	**P-value**	**Diff in CFI**	**Diff in RMSEA**	**Diff in SRMR**
M2 vs M1	112.706	2	< .001	0.008	-0.018	-0.0053
M4 vs M3	50.788	2	< .001	-0.004	0.007	0.0031

In examining the fit indexes for Models 1 and 2 (original items in a flat or hierarchical model), the flat Model 2 fit better. The chi-square difference test was statistically significant (*p* < .001), and the CFI and RMSEA and SRMR show better fit for Model 2 (Kline, 2011). In examining the fit indexes for Models 3 and 4 (alternative items in a flat or hierarchical model), the flat Model 2 fit better. Therefore in both cases, the flat structure was a better fit. Of the four models, the best fit was seen with the alternate items and supports moving forward with the alternative items that do not reference doing business or purchasing on a website.

To assess convergent validity, the factors in the SUPR-Q (using the new alternate trust items) were correlated with the Attitude toward the Website scale (Aws) from 120 participants. The correlations are shown in Table 16, and were all statistically significant at *p* < .001. The correlation between the overall SUPR-Q score and Aws Score was high (*r* = .84) suggesting good evidence for convergent validity. The subscales showed medium to strong correlations as well (*r* = .34 to *r* =.81).

Table 16

Correlations between SUPR-Q, SUPR-Q Subscales, and Aws

	Aws
SUPR-Q	.84
Usability	.77
Trust	.58
Loyalty	.75
Appearance	.81

To assess the content validity of the SUPR-Q items, the three experts in interface evaluation and user experience measurement rated how difficult they felt each of the items would be for participants to agree to. The ratings by judge, the average of the judges' ratings, and the average scores from the participants in Study 4 are shown in Table 17. The scale the judges used is inverted: lower ratings by the experts indicate items that are easier for participants to agree with. So higher scores on the SUPR-Q and lower scores by the judges indicate higher agreement.

Table 17

SUPR-Q Average Responses by Item and Experts (E1 – E3) Ratings

Item	**E 1**	**E 2**	**E 3**	**Judges' Avg.**	**Participants' Avg.**
This website is easy to use.	1	1	1	1.0	4.37
It is easy to navigate within the website.	2	1	1	1.3	4.30
The information on this website is trustworthy.	3	3	3	3.0	4.21
The information on this website is credible.	2	2	3	2.3	4.22
How likely are you to recommend this website to a friend or colleague?	1	1	1	1.0	4.15
I will likely visit this website in the future.	1	1	1	1.0	4.36
I find the website to be attractive.	3	2	2	2.3	4.17
The website has a clean and simple presentation.	3	1	2	2.0	4.27

There was good agreement between the experts. The interrater Spearman correlation between scores by each judge ranged from rho = .62 to rho = .86 (E1 vs E2 = .62; E1 vs E3 = .75; E2 vs E3 = .86). The Spearman correlation between the average expert rating and participant average was moderate, rho = -.44, *p* = .27, and not statistically significant. The *n* for this correlation was only eight. The negative correlation suggests the items that experts rated as easier for participants to agree with also tended to be the items participants agreed with more, providing support for content validity. The empirical data are based on a large sample size, and despite the small sample of experts, both are using the scale as a similar “ruler” with the judgment of the harder and easier items being in general agreement. This supports the idea that website visitors are responding in a manner generally consistent with the way experts see the construct.

Study 5

The mean scores from the two evaluators, their average score for eight of the websites, and the SUPR-Q scores and subfactors from all 16 websites are shown in Table 18. While a total of 16 websites were rated, only eight of the websites were rated by both evaluators. The average scores from the CEG are derived by averaging the responses to the 105 items (5 is the highest possible score). For example, websites with lower CEG scores were Chipotle and Bicycle Doctor USA (average scores of 1.32), and websites with higher CEG scores were Etsy and Home Depot (average scores of 4.25 and 4.17, respectively).

Table 18

Evaluator Scores from the CEG, SUPR-Q Scores and Subfactors for 16 Websites

	Eval. 1	**Eval. 2**	**Eval. Avg.**	**SUPR-Q %**	**SUPR-Q Avg.**	**Usability**	**Trust**	**Loyalty**	**Appearance**
Chipotle	1.32			0.56	3.9	3.9	4.3	3.5	3.8
eBay	3.79			0.59	3.9	3.7	4.3	3.9	3.8
Microcenter	3.51			0.54	3.9	4.2	3.9	3.3	4.0
Enterprise	2.32			0.71	4.0	4.2	4.4	3.7	3.8
UL Workplace	2.49			0.17	3.3	3.3	3.8	2.5	3.5
BicycleDoctorUS	1.32			0.01	2.0	1.9	3.0	1.5	1.5
Craigslist	3.81			0.52	3.8	4.2	3.3	4.2	3.7
Etsy	4.25			0.65	4.0	4.1	4.1	3.5	4.1
Newsweek	3.02	3.83	3.43	0.24	3.4	3.5	3.9	2.8	3.5
Wired	3.61	4.03	3.82	0.54	3.9	4.0	4.0	3.6	3.9
Macy’s	3.78	4.57	4.18	0.92	4.3	4.5	4.5	4.0	4.3
Home Depot	4.17	4.66	4.41	0.85	4.2	4.5	4.3	4.1	4.0
AirMac	2.91	3.08	3.00	0.65	4.0	4.6	4.1	3.2	4.1
Harbor Freight	3.86	3.72	3.79	0.39	3.7	4.0	4.0	3.3	3.4
Adobe	3.35	4.75	4.05	0.43	3.7	3.4	4.3	3.6	3.6
Dell	3.76	4.62	4.19	0.52	3.8	3.8	4.2	3.5	3.8

Table 19 shows the correlations between the CEG scores and SUPR-Q raw and percentile rank scores for both evaluators (and their average for eight websites) along with the corresponding SUPR-Q factor scores. The correlation in CEG scores for the two independent evaluators was moderate to high, suggesting good agreement, although with only eight observations, the correlation was not statistically significant (*r* = .61, *p* = .11).

Table 19

Correlations between CEG Scores and SUPR-Q Raw Scores, Percentile Scores and Factors

	Eval. 1	**Eval. 2**	**Avg. Eval.**
SUPR-Q Raw	.60*	.24	.58*
SUPR-Q %	.50*	.30	.49
Usability	.57*	-.25	.52*
Trust	.30	.67	.36
Loyalty	.64*	.68	.63*
Appearance	.58*	.10	.58*

* *p* < .05

Table 19 shows that the correlation between evaluator 1 and the SUPR-Q percentile raw score was reasonably high (*r* = .60) and to a lesser extent the correlation with the percentile rank scores (*r* = .50). Evaluator 2 had much lower correlations (*r* = .30, *p* > .4) with the overall SUPR-Q raw and percentile scores, although this was computed on only eight of the websites, thus *n* was small. The correlation between the average score of the evaluators and the SUPR-Q raw score was similar to evaluator 1, with the correlation being .58 and .49 for SUPR-Q raw and percentile scores, respectively (the latter was not statistically significant).

Rasch analysis.

Data regarding the CEG were obtained from two sources: evaluators with experience examining interfaces (experts) and a selection of the general internet population recruited from Mechanical Turk (non-experts). First, data were examined using Winsteps, which ignores the potential effects of the judges on the items. The same dataset was also examined using Facets to understand if accounting for the individual judge and training of the judge affect the model fit and indices of measure quality.

A total of 225 independent evaluators rated 68 websites using randomly assigned sections of the CEG. The websites and number of evaluators assigned to each website are shown in Appendix H. Each evaluator was presented with at least one section, the section concerning navigation elements, and at least one additional section of the CEG. This consisted of at least 11 items per evaluator. Appendix F contains the names of the website and the total number of evaluators that contributed (the same dataset was used for the Winsteps and Facets analysis). Not all items were applicable to every website, or participants could not make a judgment about the compliance of the website with the item. This was particularly the case for items that dealt with purchasing, which was not applicable for most government websites. For example, item 63, "Cross selling is used appropriately; not intrusive or easily confused with cart contents," was often not applicable. Of the 105 items, 48 items were used by the evaluators at least 100 times across the 68 websites and were included in the Rasch analysis.

The scale use was examined. All five categories of the response options were sufficiently used (greater than five observations). The response scale step calibration went from low (-1.99) to high (2.39), with no reversals in order as shown below in Figure 6.

```
SUMMARY OF CATEGORY STRUCTURE.  Model="R"
-------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT|| ANDRICH |CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||THRESHOLD| MEASURE|
|-------------------+------------+------------++---------+--------|
|  1   1      297   4|  -.04  -.30|  1.35  1.86||  NONE    |( -1.99)| 1
|  2   2      408   5|   .11   .14|   .94   .98||    -.40 |   -.85 | 2
|  3   3      899  11|   .56   .58|   .99  1.11||    -.43 |   -.13 | 3
|  4   4     2737  34|   .99  1.07|   .91   .81||    -.30 |    .74 | 4
|  5   5     3749  46|  1.85  1.80|   .99   .98||    1.13 |(  2.39)| 5
|-------------------+------------+------------++---------+--------|
|MISSING    7366  48|  1.31       |            ||         |        |
-------------------------------------------------------------------
OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

--------------------------
|CATEGORY    STRUCTURE   |
|------------------------+
|   1      NONE
|   2        -.40    .07 |
|   3        -.43    .05 |
|   4        -.30    .03 |
|   5        1.13    .03 |
--------------------------
```

Figure 6. Rating scale use.

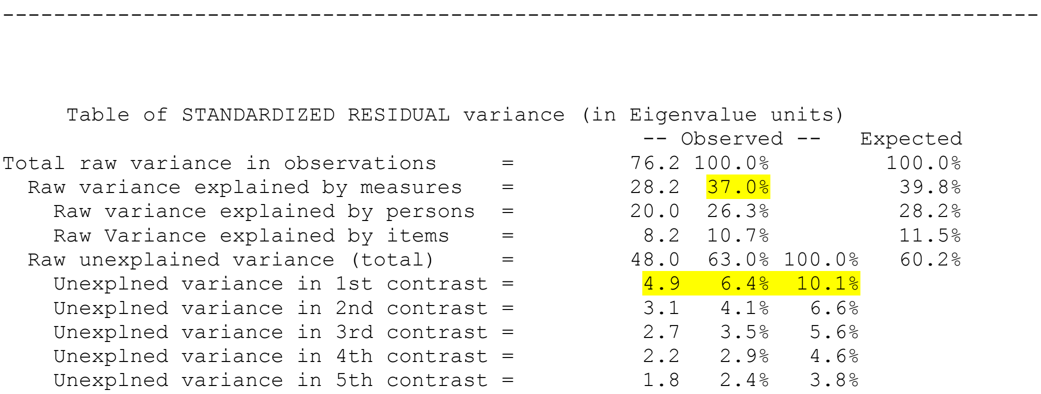
The category probability figure shown in Figure 7 displays values progressing from low to high. However, there is some evidence that respondents are not using all points. Points 1, 2, and 3 were not used extensively. There may be a need to collapse the categories or re-examine after removing poorly fitting items.

```
        CATEGORY PROBABILITIES: MODES - Structure measures at intersections

P      -+-----------+-----------+-----------+-----------+-----------+-
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |                                                         5555|
B   .8 +11                                                    555    +
I      |  11                                               555       |
L      |    11                                          555          |
I      |      11                                      55             |
T   .6 +        11                                  55               +
Y      |          11                              55                 |
    .5 +            1                           55                   +
O      |             1              4444444444*5                     |
F   .4 +              11          44        55 4444                  +
       |                1       44        55       444               |
R      |                 11   44        55            444            |
E      |       22222222222****3333333 55                 4444        |
S   .2 +  22222       3333441*22    5*33                     4444    +
P      |22         333 444    112**5    3333                     4444|
O      |      33333 444       55*112222     33333                    |
N      |333333 44444     55555     1111*22222    333333333           |
S   .0 +*******5555555555               11111************************+
E      -+-----------+-----------+-----------+-----------+-----------+-
       -2          -1          0           1           2           3
        PERSON [MINUS] ITEM MEASURE
```

Figure 7. Category probability curves for responses to the CEG.

To assess dimensionality, the table of standardized residuals is shown in Figure 8 below. There is poor evidence to support unidimensionality, since the variance explained by the measure was 37 percent, which is slightly below a common threshold of 40 percent. The unexplained variance in the first contrast, however, was above the 2.0 threshold (4.9) and higher than the recommended 5 percent (6.4 percent) (Linacre,1989).



```
--------------------------------------------------------------------------------


     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                 -- Observed --   Expected
Total raw variance in observations     =         76.2 100.0%         100.0%
  Raw variance explained by measures   =         28.2  37.0%          39.8%
    Raw variance explained by persons  =         20.0  26.3%          28.2%
    Raw Variance explained by items    =          8.2  10.7%          11.5%
  Raw unexplained variance (total)     =         48.0  63.0% 100.0%   60.2%
    Unexplned variance in 1st contrast =          4.9   6.4%  10.1%
    Unexplned variance in 2nd contrast =          3.1   4.1%   6.6%
    Unexplned variance in 3rd contrast =          2.7   3.5%   5.6%
    Unexplned variance in 4th contrast =          2.2   2.9%   4.6%
    Unexplned variance in 5th contrast =          1.8   2.4%   3.8%
```

Figure 8. Dimensionality indices.

Item fit was examined iteratively and items were removed and the model rerun until only the items with infit/outfit values above 1.4 were retained (Linacre, 1989). This left 23 items, which are indicated in Table 20 (relative to the 48) and in Appendix D (relative to the original 105). An examination of the removed items shows they are associated with search and the general design and layout. The bulk of the retained items relate to content and navigation.

Table 20

Retained 18 Items Relative to the 48 Used Initially in Winsteps

Item Code	**Item Description**	**Retained = 1**
PI1	Content is concise	1
PI2	Content appropriate form	1
PI3	Up to date info	0
PI4	Logical layout	1

PI5	Easy to scan	1
PI6	Content is easy to find	1
PI7	No distractions	1
PI8	Easy to find info	1
PI9	Print/Download	0
PI10	Effective formatting	0
PI11	Content is usable	1
N1	Nav is logical	1
N2	Nav is direct	1
N3	Aware of location	1
N4	Organized/navigable	1
N5	Deviate from path	0
N6	Button labels	1
N7	Nav is visible	1
N8	Scrolling	0
N9	Link destinations	0
N10	Pace of nav	1
N11	Nav is usable	1
S1	Search is intuitive	0
S2	Edit search terms	0
S3	Clear search results	0
S4	Search filter	0
S5	Search misspellings	0
S6	Alternate spelling	0
S7	Advanced search exists	0
S8	Scope of search	0
S9	Search result views	0
S10	Predictive search	0
S11	Search is usable	0
G1	Experience	1
G2	Site functioning correctly	0
G3	Includes logical pages	1
G4	Ads and popups	0
G6	Screen space	1
G7	Novel devices	0
G9	Site is readable	1
G10	Color contrast	0
G11	Text colors readable	1
G12	Layout is balanced	0
G13	Visual design	0

G14	Design is relevant	1
G15	Design isn’t flashy	0
G16	Original design	0
G17	Overall usable	1

The person fit was next examined, and only three judges had infit or outfit mean squares greater than 3 as shown in Figure 9 and were removed (Bond & Fox, 2007).

```
-----------------------------------------------------------------------------------------
--------------------------------------------------
|ENTRY   TOTAL  TOTAL          MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT MATCH|
|
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| PERSON
|
|------------------------------------+----------+----------+-----------+-----------+-------------
-------------------------------------------------|
|   138     83     23     .53     .27|5.02   6.6|4.83   6.3|A .72   .28|   .0  51.5| 143 NonMaster
|    50     53     12    2.21     .50|3.86   4.7|4.08   5.0|B-.18   .11| 33.3  50.9|  50 NonMaster
|   318     57     15     .90     .36|3.31   3.5|3.00   3.2|C .21   .27| 26.7  59.6| 353 Master
|   262     63     16    1.21     .37|2.71   2.9|2.64   2.8|D .55   .25| 18.8  60.8| 287 Master
|   177     71     15    3.37     .57|2.69   3.2|2.26   2.6|E .62   .08| 86.7  73.9| 187 NonMaster
|   322     31      8     .77     .51|2.49   1.9|2.55   2.0|F-.65   .08| 25.0  62.7| 358 NonMaster
|   279     34     15   -1.26     .27|2.46   3.7|2.44   3.7|G .00   .17| 13.3  36.8| 306 Master
|   167     91     23    1.19     .31|2.34   2.9|2.39   3.0|H .18   .24| 26.1  61.5| 175 NonMaster
|    91    107     23    3.21     .42|2.19   3.4|1.94   2.8|I .34   .18| 73.9  66.8|  96 NonMaster
|    57     77     16    4.08     .63|1.85   1.6|2.17   1.9|J-.11   .16| 87.5  81.5|  59 NonMaster
```

Figure 9. Winsteps person fit.

After removing the three judges, the dimensionality improved. The raw variance was above 40 percent (now at 52.5 percent) and the first contrast unexplained variance was slightly above 2 (2.2) but below 5 percent, indicating reasonable evidence for unidimensionality as shown in Figure 10.

```
------------------------------------------------------------------------------
-----

     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                 -- Observed --   Expected
Total raw variance in observations     =        48.4 100.0%         100.0%
  Raw variance explained by measures   =        25.4  52.5%          52.5%
    Raw variance explained by persons  =        23.7  48.9%          48.9%
    Raw Variance explained by items    =         1.7   3.5%           3.5%
  Raw unexplained variance (total)     =        23.0  47.5% 100.0%   47.5%
    Unexplned variance in 1st contrast =         2.2   4.6%   9.6%
    Unexplned variance in 2nd contrast =         2.1   4.3%   9.1%
    Unexplned variance in 3rd contrast =         1.5   3.2%   6.7%
    Unexplned variance in 4th contrast =         1.5   3.0%   6.3%
    Unexplned variance in 5th contrast =         1.4   2.9%   6.1%
```

Figure 10. Winsteps dimensionality output.

The scale use was re-examined. All five categories were sufficiently used (greater than five observations), although 87 percent of the observations are scored 4 or 5. The response scale step calibration went from low (-3.21) to high (3.87) as shown in Figure 11.

```
-------------------------------------------------------------------
------

SUMMARY OF CATEGORY STRUCTURE.  Model="R"
-------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT|| ANDRICH |CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||THRESHOLD| MEASURE|
|-------------------+------------+------------++---------+--------|
|  1   1      42   1| -1.34 -1.36|  1.04  1.11||  NONE   |( -3.21)|
1
|  2   2     139   3|  -.21  -.19|   .98  1.02||   -1.92 |  -1.53 |
2
|  3   3     391   9|   .85   .78|  1.08  1.23||    -.74 |   -.29 |
3
|  4   4    1601  37|  1.94  1.98|   .92   .90||    -.06 |   1.44 |
4
|  5   5    2144  50|  3.59  3.56|  1.01   .98||    2.73 |(  3.87)|
5
|-------------------+------------+------------++---------+--------|
|MISSING    2836  40|  2.68      |            ||         |        |
-------------------------------------------------------------------
OBSERVED AVERAGE is mean of measures in category. It is not a
parameter estimate.
```

Figure 11. Scale use for items and persons retained.

The category probability map displays values going from low to high as shown in Figure 12. However, there is some evidence that respondents are not using all points. Point 3 to some extent is not being used as expected. There may be a need to collapse the categories in future work.

```
        CATEGORY PROBABILITIES: MODES - Structure measures at
intersections
P      -+------+------+------+------+------+------+------+------+-
R  1.0 +                                                         +
O      |                                                         |
B      |1                                                        |
A      | 111                                                     |
B   .8 +    11                                                  5+
I      |      11                                               5 |
L      |        1                                            55  |
I      |         1                         4444444         55    |
T   .6 +          11                     44       44      5      +
Y      |            1                   4           44   5       |
    .5 +             1                44              4*5        +
O      |              122222         4                5 4        |
F   .4 +             221    222  3334                5   44      +
       |           22   1     3*3  4333            55      44    |
R      |         22      1  33  224    33         5          4   |
E      |       22         1*     42      33     55            44 |
S   .2 +    222          33 1  44  22      33 55                4+
P      | 222           33    1*      2      5*33                 |
O      |2           333     44 11     222555    333              |
N      |       33333    4444     111*55552222      33333         |
S   .0 +****************555555555555 11111111********************+
E      -+------+------+------+------+------+------+------+------+-
       -4     -3     -2     -1      0      1      2      3      4
        PERSON [MINUS] ITEM MEASURE
```

Figure 12. Scale usage for retained items.

A further examination of the 23 items shows that the items tend not to be difficult enough, as there are many websites scoring higher than the highest items as shown in the clustering in Exhibit 13. This suggests that future items should be considered that are both harder and easier to agree to.

```
TABLE 1.12 topitemsPass5.xls                     ZOU055WS.TXT  Aug 14 12:16 2015
INPUT: 324 PERSON  23 ITEM  REPORTED: 311 PERSON  23 ITEM  5 CATS WINSTEPS 3.81.0
---------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
               <more>||<frequent>
    8   .##########  ++
                     ||
                     ||
                     ||
                     ||
    7                ++
                     ||
                     ||
                     ||
                     ||
    6                ++
                  .  ||
                  . T||
                  .  ||
                 ##  ||T
    5             .  ++  Button  Design  Includ
                     ||S Aware   Conten  Nav is  Nav is  Overal
                .##  ||M Conten  Nav is  Nav is  Organi  Pace o  Site i  Text c
                .##  ||  Easy t  Experi
                  .  ||S Conten  Conten  Logica  Screen
    4           .## S++  Easy t
                 .#  ||T
               .###  ||  No Dis
                 .#  ||
               .###  ||
    3           ###  ++
                 .#  ||
               .###  ||
               .###  ||
                .## M||
    2           .##  ++
                 .#  ||
                ###  ||
                .##  ||
              #####  ||
    1           ###  ++
                 .#  ||
               .### S||
                  #  ||
                .##  ||
    0            .#  ++
                  #  ||
                  .  ||
                  .  ||
                  .  ||
   -1             . T++
                     ||
                  .  ||
                     ||
                     ||
   -2                ++
                     ||
                     ||
                     ||
                     ||
   -3                ++
                     ||
                  .  ||
                     ||
                     ||
   -4                ++
               <less>||<rare>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

Figure 13. Item map.

Many-faceted analysis of the CEG. The analysis done in Winsteps ignored the effects of the judges. To understand the effect of judge (research question 6), judge's experience level, and website on rating of the items in the CEG, the dataset was analyzed using Facets 3.71.4 (Linacre, 2015). Only the items with at least 100 responses were used, leaving data from 48 items, 68 websites, and 225 judges. The initial pass on Facets identified 14 judges with infit mean square values in excess of 3.0 or that fell below a minimum threshold for computation (i.e., an insufficient set of ratings)—both suggested poor model fit. These judges were removed and the model rerun. In the second pass, all judges had infit mean square values less than 3.0.

The table of items was examined; six items had infit mean square values greater than 1.5 and were removed. After running the model again (pass three), five new judges were flagged for having infit mean-square values in excess of 3 and were removed. On the fourth pass, all judges had infit mean-square values less than 3, but five items had infit mean-square values greater than 1.5 and were removed. The fifth pass identified two judges with infit mean squares greater than 3 and were removed for a sixth pass. After the sixth pass, 202 judges remained. The experience and website facets showed good fit in all analyses.

An analysis of the scale use echoed what was found with the Winsteps (single facet model), with the lower ends of the response scale being sparsely used. Only 1 percent of judges used response option 1, 4 percent used response option 2, and 11 percent used response option 3.

Figure 14 below shows the probability values for each scale step progressing from low to high. However, points 2 and 3 were not used extensively. This suggests that judges were not fully using the five points in the scale, and the number of response options were then collapsed (similar to the findings from Winsteps).

```
-4.0             -2.0              0.0              2.0              4.0
   ++----------------+----------------+----------------+----------------++
 1 |                                                                     |
   |1                                                                    |
   | 11111                                                             55|
   |      111                                                       555  |
   |         11                                                  555     |
 P |           11                                              55        |
 r |             11                                          55          |
 o |               11                                      55            |
 b |                 1                                    5              |
 a |                  11                      4444      55               |
 b |                    1                  444    444  5                 |
 i |                     1               44          **4                 |
 l |                     2*22222       44           5   44               |
 i |                  222  11   222333*           55      44             |
 t |               222       1  3332 4 3333      5          44           |
 y |             22           *3   4*2     33  55             444        |
   |          222           33 11 4   22     **3                 44      |
   |      2222           333    4*1     22 55   33                 4444  |
   | 22222            333    444   11   55*22     3333                 44|
   |2           333333   4444      55***1    2222     333333             |
 0 |*********************5555555555      11111111************************|
   ++----------------+----------------+----------------+----------------++
  -4.0             -2.0              0.0              2.0              4.0
```

Figure 14. Probability curves for many-faceted CEG model.

The response scale was then collapsed into three points, with points 1, 2, and 3 being consolidated, and 4 and 5 coded as 2 and 3, respectively. The analysis was rerun; two judges had maximum scores, and one site had a minimum score. Both were removed and the analysis was run again.

On the eighth pass, an examination of the judges, items, experience, and websites all showed good fit. The item separation and reliability are shown in Figure 15, and the 37 retained items are shown in Appendix G.

```
+-----------------------------------------------------------------------------------------------------
---------
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Correlation |
|
|
|-------------------------------+--------------+---------------------+------+-------------+---------
------------|
|   294     117      2.51   2.54 |   -.73    .18 | 1.17  1.2  1.29  1.1 |  .75 |   .46   .57 |  3 PI3
|   403     167      2.41   2.47 |   -.52    .14 |  .85 -1.4   .85  -.3 | 1.18 |   .63   .58 | 26 G3
|   645     266      2.42   2.46 |   -.49    .11 | 1.00   .0  1.13   .5 |  .94 |   .53   .56 | 17 N6
|   306     123      2.49   2.46 |   -.49    .17 |  .89  -.8   .75  -.5 | 1.22 |   .59   .52 | 34 G14
|   634     264      2.40   2.43 |   -.40    .11 |  .72 -3.6   .69 -1.2 | 1.36 |   .66   .56 | 21 N11
|   281     117      2.40   2.40 |   -.34    .17 |  .60 -3.6   .56 -2.4 | 1.49 |   .74   .59 | 11 PI11
|   281     117      2.40   2.39 |   -.31    .17 |  .89  -.8   .84  -.7 | 1.19 |   .65   .58 |  1 PI1
|   617     261      2.36   2.39 |   -.30    .11 |  .92  -.9   .87  -.4 | 1.12 |   .59   .57 | 18 N7
|   386     165      2.34   2.38 |   -.27    .14 |  .53 -5.3   .55 -1.7 | 1.54 |   .74   .60 | 37 G17
|   624     265      2.35   2.37 |   -.26    .11 |  .95  -.6   .94  -.1 | 1.04 |   .57   .57 | 14 N3
|   617     262      2.35   2.37 |   -.26    .11 |  .69 -4.2   .64 -1.5 | 1.40 |   .67   .57 | 15 N4
|   619     263      2.35   2.37 |   -.25    .11 |  .79 -2.7   .76  -.9 | 1.30 |   .65   .57 | 13 N2
|   297     123      2.41   2.36 |   -.24    .16 |  .73 -2.4   .73  -.6 | 1.34 |   .64   .54 | 29 G9
|   297     123      2.41   2.36 |   -.24    .16 |  .98  -.1  1.16   .5 | 1.02 |   .55   .54 | 32 G12
|   294     122      2.41   2.35 |   -.21    .16 | 1.08   .6  1.07   .3 |  .97 |   .55   .54 | 31 G11
|   610     264      2.31   2.31 |   -.11    .11 |  .86 -1.8   .91  -.3 | 1.11 |   .58   .57 | 20 N10
|   608     265      2.29   2.30 |   -.08    .11 |  .76 -3.2   .70 -1.4 | 1.29 |   .64   .58 | 12 N1
|   377     166      2.27   2.30 |   -.07    .14 | 1.32  2.8  1.43  1.5 |  .61 |   .52   .61 | 28 G7
|   256     113      2.27   2.29 |   -.06    .16 | 1.70  4.7  1.81  3.7 |  .07 |   .38   .61 | 22 S1
|   272     118      2.31   2.28 |   -.03    .16 |  .86 -1.1   .78 -1.2 | 1.24 |   .67   .60 |  8 PI8
|   285     122      2.34   2.27 |    .00    .16 | 1.09   .7  1.06   .2 |  .98 |   .57   .56 | 30 G10
|   270     118      2.29   2.26 |    .02    .16 |  .94  -.4   .91  -.4 | 1.10 |   .63   .60 |  6 PI6
|   249     112      2.22   2.23 |    .08    .16 | 1.33  2.4  1.46  2.3 |  .62 |   .53   .61 | 24 S11
|   250     113      2.21   2.22 |    .10    .16 | 1.32  2.4  1.47  2.4 |  .51 |   .47   .61 | 23 S8
|   367     166      2.21   2.22 |    .11    .14 |  .95  -.4  1.06   .3 |  .99 |   .60   .61 | 25 G1
|   266     118      2.25   2.21 |    .13    .16 | 1.11   .9  1.00   .0 |  .89 |   .58   .61 |  2 PI2
|   266     118      2.25   2.21 |    .13    .16 |  .95  -.3   .86  -.7 | 1.10 |   .62   .61 |  4 PI4
|   282     123      2.29   2.21 |    .14    .16 | 1.26  2.1  1.19   .6 |  .77 |   .55   .57 | 35 G15
|   367     167      2.20   2.20 |    .16    .13 |  .92  -.8   .94  -.1 | 1.10 |   .64   .62 | 27 G6
|   262     117      2.24   2.19 |    .17    .16 | 1.02   .2   .95  -.2 | 1.03 |   .62   .61 |  5 PI5
|   274     122      2.25   2.15 |    .27    .16 | 1.18  1.5  2.10  2.9 |  .62 |   .49   .58 | 33 G13
|   255     118      2.16   2.09 |    .40    .16 | 1.18  1.5  1.35  2.0 |  .79 |   .59   .61 | 10 PI10
|   267     123      2.17   2.05 |    .50    .15 | 1.40  3.1  1.61  2.0 |  .38 |   .46   .59 | 36 G16
|   537     262      2.05   1.98 |    .64    .10 | 1.12  1.4  1.14   .8 |  .77 |   .52   .60 | 16 N5
|   245     118      2.08   1.98 |    .64    .16 | 1.06   .5   .97  -.1 | 1.03 |   .64   .61 |  7 PI7
|   502     263      1.91   1.81 |   1.04    .10 | 1.44  4.9  1.47  2.8 |  .38 |   .48   .60 | 19 N8
|   203     106      1.92   1.77 |   1.13    .16 | 1.58  4.1  1.66  3.5 |  .15 |   .43   .61 |  9 PI9
|-------------------------------+--------------+---------------------+------+-------------+---------
---------
|   374.7   164.0    2.28   2.26 |    .00    .14 | 1.03   .0  1.07   .3 |      |   .58       | Mean
(Count: 37)
|   145.4    62.7     .14    .17 |    .40    .02 |  .26  2.4   .36  1.5 |      |   .08       | S.D.
(Population)
|   147.4    63.6     .14    .17 |    .41    .02 |  .26  2.5   .36  1.5 |      |   .08       | S.D.
(Sample)
+-----------------------------------------------------------------------------------------------------
---------
Model, Sample: RMSE .15  Adj (True) S.D. .38  Separation 2.61  Strata 3.81  Reliability .87
Model, Fixed (all same) chi-square:  341.8  d.f.: 36  significance (probability): .00
Model,  Random (normal) chi-square:  32.1  d.f.: 35  significance (probability): .61
-----------------------------------------------------------------------------------------------------
---------
```

Figure 15. Item measurement report (arranged by MN).

Detailed Facets tables are shown in Appendix G. The probability curves for the new three-point scale are shown in Figure 16 and show a pattern of increasing probability for the three points.

```
      -3.0       -2.0       -1.0        0.0        1.0        2.0        3.0
       ++----------+----------+----------+----------+----------+----------++
     1 |                                                                   |
       |                                                                   |
       |11                                                               33|
       |  1111                                                       3333  |
       |      1111                                               3333      |
     P |          11                                           33          |
     r |            111                                     333            |
     o |               11                                 33               |
     b |                 11                             33                 |
     a |                   11        222222222        33                   |
     b |                     11  2222         2222  33                     |
     i |                      2**                 **2                      |
     l |                   222   11             33   222                   |
     i |                222        11         33        222                |
     t |             222             11     33             222             |
     y |          222                  11*33                  222          |
       |       222                     33 11                     222       |
       |  22222                     333     111                     22222  |
       |22                     33333           11111                     22|
       |               33333333                     11111111               |
     0 |333333333333333                                     111111111111111|
       ++----------+----------+----------+----------+----------+----------++
                                                                                                     -
3.0       -2.0       -1.0        0.0        1.0        2.0        3.0
```

Figure 16. Probability curves for collapsed three-point scale.

Figure 17 shows the “rulers,” which show little separation between master and non-master but good separation between master/nonmaster and expert. There is a good spread in website representation; however, the items are clustered within a narrow range (within +1/-1 logit position).

```
Vertical = (1*,2A,3*,4*,S) Yardstick (columns lines low high extreme)= 0,3,-7,9,End
+------------------------------------------------------------------+
|Measr|-judge     |-experience          |+site     |-item    |Scale|
|-----+-----------+---------------------+----------+---------+-----|
|   9 +           +                     +          +         + (3) |
|     |           |                     |          |         |     |
|     | *.        |                     |          |         |     |
|   8 +           +                     +          +         +     |
|     |           |                     |          |         |     |
|     | *         |                     |          |         |     |
|   7 + .         +                     + *        +         +     |
|     | *.        |                     |          |         |     |
|     |           |                     | *        |         |     |
|   6 + *.        +                     + *        +         +     |
|     | .         |                     | **       |         |     |
|     |           |                     |          |         |     |
|   5 + *.        +                     + *****    +         +     |
|     | .         |                     | **       |         |     |
|     | .         |                     | **       |         |     |
|   4 + **        +                     + **       +         +     |
|     | .         |                     | *        |         |     |
|     | **.       |                     | *        |         |     |
|   3 + *.        +                     +          +         +     |
|     | .         |                     | **       |         |     |
|     | .         |                     | *        |         |     |
|   2 + **.       + expert              +          +         +     |
|     | **.       |                     | *        |         |     |
|     | ****.     |                     |          |         | --- |
|   1 + ***.      +                     +          + *       +     |
|     | ******    |                     | *        | *.      |     |
|     | *****     |                     | **       | *.      |     |
*   0 * ***       *                     * ****     * ******* *  2  *
|     | ******    |                     | ******** | ******. |     |
|     | *****     |                     | ****     | *       |     |
|  -1 + ********. + master    nonmaster + ***      +         +     |
|     | *******   |                     | ****     |         | --- |
|     | *****     |                     | ****     |         |     |
|  -2 + *****     +                     + ******   +         +     |
|     | *****     |                     | *        |         |     |
|     | **.       |                     | ***      |         |     |
|  -3 + ***.      +                     + *        +         +     |
|     | *.        |                     | *        |         |     |
|     | .         |                     |          |         |     |
|  -4 + ***.      +                     +          +         +     |
|     | *         |                     | *        |         |     |
|     | .         |                     |          |         |     |
|  -5 + .         +                     + *        +         +     |
|     | .         |                     |          |         |     |
|     |           |                     |          |         |     |
|  -6 +           +                     +          +         +     |
|     | .         |                     |          |         |     |
|     |           |                     |          |         |     |
|  -7 +           +                     +          +         + (1) |
|-----+-----------+---------------------+----------+---------+-----|
|Measr| * = 2     |-experience          | * = 1    | * = 2   |Scale|
```

Figure 17. All Facet vertical “rulers.”

Figure 18 shows a closer view of the distribution of the items by item position, with clusters of items between -.27 and -.24 logits, and between .13 and .17 logits.

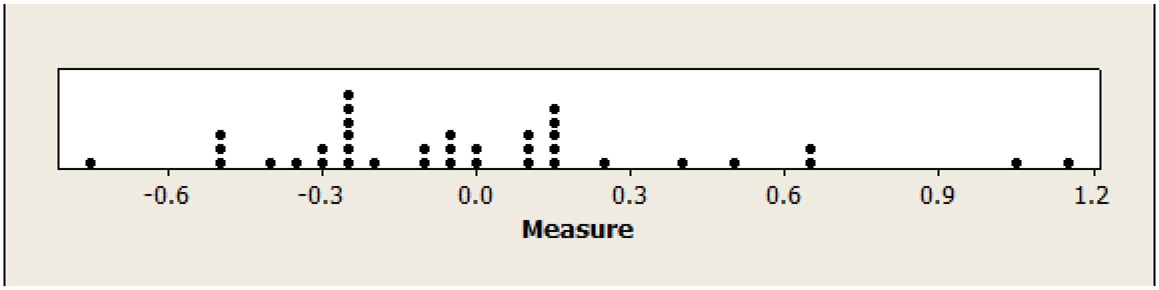


Figure 18. Item distribution for the 37 items in the CEG.

While the differences among items were significant but not extensive, there were significant effects on ratings of both judge expertise and of individual judges. As seen in Figures 19 and 20, the test of difference in rating by judge's expertise was statistically significant, and the test of difference in rating based on the individual judge was statistically significant. The test for significance of difference is the model, fixed chi-square, that tests whether all elements of the facet take an equivalent position on the logit scale (highlighted in the respective tables). The fact that judges differed significantly suggests that judge training is needed, and potentially, multiple judges are needed to provide a good rating of websites.

```
Model, Populn: RMSE .47  Adj (True) S.D. 2.78  Separation 5.88  Strata 8.17  Reliability (not inter-

rater) .97


Model, Sample: RMSE .47  Adj (True) S.D. 2.79  Separation 5.89  Strata 8.19  Reliability (not inter-
rater) .97
Model, Fixed (all same) chi-square:  9332.8  d.f.: 198  significance (probability): .00
Model,  Random (normal) chi-square:  190.6  d.f.: 197  significance (probability): .62
Inter-Rater agreement opportunities: 6663  Exact agreements: 2766 =  41.5%  Expected:  2699.1 =  40.5%
------------------------------------------------------------------------------------------------------
---------Note: Full table output is in Appendix G.
```

Figure 19. Selection of output for Judge Measurement Report.

```
+-----------------------------------------------------------------------------------------------------

------------+


|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Correlation |
|
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | N
experience        |
|--------------------------------+--------------+----------------------+------+-------------+---------
------------|
|  9620    4135      2.33   2.64 |  -1.05   .03 |  .98  -.8  1.01   .1 | 1.03 |   .60   .59 | nonmaste
|  2951    1299      2.27   2.64 |  -1.03   .05 | 1.01   .4  1.01   .1 | 1.00 |   .55   .55 | 2 master
|  1294     633      2.04   1.41 |   2.09   .07 | 1.14  2.5  1.23  3.3 |  .81 |   .65   .69 | 1 expert
|--------------------------------+--------------+----------------------+------+-------------+---------
|  4621.7  2022.3    2.21   2.23 |    .00   .05 | 1.05   .7  1.08  1.2 |      |   .60       | Mean
|  3598.5  1518.4     .12    .58 |   1.47   .02 |  .07  1.4   .10  1.5 |      |   .04       | S.D.
|  4407.3  1859.7     .15    .71 |   1.81   .02 |  .08  1.7   .13  1.9 |      |   .05       | S.D.
+-----------------------------------------------------------------------------------------------------
------------+
Model, Populn: RMSE .05  Adj (True) S.D. 1.47  Separation 28.79  Strata 38.73  Reliability 1.00
Model, Sample: RMSE .05  Adj (True) S.D. 1.81  Separation 35.27  Strata 47.36  Reliability 1.00
Model, Fixed (all same) chi-square:  1795.3  d.f.: 2  significance (probability): .00
Model,  Random (normal) chi-square:  2.0  d.f.: 1  significance (probability): .16
-------------------------------------------------------------------------------------------------
```

Figure 20. Judge's Experience Measurement Report (arranged by MN).

For future research, a more compact version of the CEG can be created by removing items with similar positions. Items that are candidates for removal are the ones close in position to another item (within .01 points of another item). Items with the higher number of respondents and lower mean-square infit value were retained for identical or adjacent items. This left 18 items as shown in Table 19. Appendix D also lists items that were also identified using Winsteps, with 10 overlapping between the two models. The names of the websites used in the analysis are also included in Appendix H.

Table 21

Logit Positions by Item Retained

Item	**Measure**	**Infit MSq**	**N**	**Label**
PI3	-0.73	1.17	117	The information is recent and up to date.
G3	-0.52	0.85	167	The website contains a logical selection of relevant pages, e.g., home page, about, products.
N11	-0.4	0.72	264	The website's navigation is usable for the typical user.
N7	-0.3	0.92	261	Navigational options are visible and obvious.
G17	-0.27	0.53	165	The website overall is usable for the typical user.
G11	-0.21	1.08	122	Text is in colors and fonts that are readable and appropriate for the site.
N10	-0.11	0.86	264	The pace of the navigation does not rush the user or hold them back from accessing content.
N1	-0.08	0.76	265	Navigation options are ordered in a logical or task-oriented manner.
PI8	-0.03	0.86	118	It is easy to locate and identify desired information.
G10	0	1.09	122	The text and background are contrasting colors.
PI6	0.02	0.94	118	Content is easy to find and get to.
G6	0.16	0.92	167	The website fully utilizes its resources, e.g., screen space, navigation, etc.
G13	0.27	1.18	122	The visual design is appealing to the typical user.
PI10	0.4	1.18	118	Content formatting is effective and visually appealing.
G16	0.5	1.4	123	The visual design is original and related to the website's other branding.
N5	0.64	1.12	262	It is easy to leave or deviate from a navigational path, but it is clear when a user is doing so.
N8	1.04	1.44	263	There is no unnecessary scrolling or panning.
PI9	1.13	1.58	106	It is easy to save, print, or download information as appropriate.

CHAPTER 4: DISCUSSION

This study used classical test theory, confirmatory factor analysis, and two Rasch models to develop two new instruments to measure the quality of the website user experience.

The first measure, the SUPR-Q, is intended to be administered to website users; it is brief and backed by a normative database. The second instrument, the Calibrated Evaluator’s Guide to User Experience Quality (CEG), is intended for use in evaluating elements of a website by independent judges based on guidelines that provide more diagnostic and detailed information on what to improve. Existing measures are either not comprehensive enough, do not have a normative database (or are proprietary), are too long, or do not address the diagnostic needs.

This chapter provides a discussion of the findings presented in Chapter 3 as they relate to the original research questions. Finally, research limitations and recommendations for future research are discussed.

Research question 1

Research question 1 sought to identify the aspects that best quantify the quality of the website user experience. The literature review revealed that the most common constructs related to measuring the quality of the website user experience are usability, trust, appearance, and loyalty. The literature also suggested that these constructs were correlated. In Study 1, concepts were garnered from the literature and items written to represent the constructs of usability, trust, appearance, and loyalty. The results of the factor analysis in Study 1 were not surprising. It is often the case that “you get out what you put in” (items to represent four constructs in and four factors out, but only if you have identified four reasonably independent constructs and have done a good job of selecting items to measure them). However, the factor analysis identified which of the original items loaded highest on the retained factors, had lowest cross-loadings on other factors, had a strong item–total correlation, and contributed to Cronbach’s alpha. So while it is very common to expect a certain factor structure, it was unclear until data were collected in Study 1 which set of items, if any, would have the desired attributes.

Research question 2

Research question 2 sought to identify the items that had the best psychometric properties. The items in Study 1 were found to correlate highly with the System Usability Scale providing encouraging evidence of convergent validity—even with a small sample and preliminary items. The items also exhibited a high Cronbach’s alpha.

Study 2 added additional data and a larger sample size to examine the psychometric properties of the items identified in Study 1. The four-factor structure was replicated and convergent validity was established with the SUS ($r \geq .87$) and WAMMI ($r \geq .85$) instruments. The new set of reduced items and subfactors also exhibited high internal consistency reliability, with the exception of the loyalty factor. The loyalty factor had reliability Cronbach’s alpha that was lower than desired (alpha = .63). The widespread usage of the Net Promoter Score means that the lower reliability is offset by its comparability.

In Study 4, additional items for the trust factor were examined to account for websites that had no e-commerce component (e.g., informational and government websites). The confirmatory factor analysis (CFA) found the eight items fit the four-factor model well. The new items (“The information on the website is trustworthy” and “The information on the website is credible”) were found to actually be better representations of the trust factor—showing better model fit on the flat (non-hierarchical) model than the items “I feel comfortable purchasing from the website” and “The website keeps the promises it makes to me.”

Research question 3

Research question 3 asked if the new SUPR-Q instrument has sufficient reliability while remaining short.

Study 3 further refined the items down to just eight (two per factor), the minimum number possible to still perform a reliability analysis while building a normative database of comparable websites. There was strong evidence for convergent validity with the SUS and WAMMI, and the SUPR-Q was able to discriminate as well as those other instruments among 40 websites. The overall composite score and Usability factors both had high reliability (coefficient alpha >.85), and the trust and appearance factors had acceptable reliability (coefficient alpha >.75). However, the loyalty factor again showed lower reliability (coefficient alpha=.64), as was found in Study 2.

There are two possible explanations for the low internal consistency reliability. First, the loyalty factor uses an 11-point scale to keep the scoring consistent with industry practices (specifically the Net Promoter Score). It is possible that having different numbers of response options affects the correlation and thus estimates of internal consistency reliability. The loyalty factor has only two items, and the tradeoff for fewer items is lower reliability, which in this case, while low, is similar to that found for factors on the WAMMI (which is more than twice as long, at 20 items). Second, it is likely that low prior experience with the websites may play a role in how likely participants are to recommend the website. For several of the websites used in the sample, participants had low to no prior experience with the website before they attempted tasks. Future research can examine whether prior experience has an effect on the reliability of the loyalty factor. Furthermore, the "likelihood to recommend" item had the lowest average rating from Study 4 compared to the other seven items and also corresponded to the lowest rating from the three experts—again suggesting less willingness to recommend. It is plausible that this less willingness to recommend is a function of prior experience with the websites. If participants have little or no experience with the website, a question asking about recommending to friends may become less appropriate and hence contribute to lower reliability.

Despite the shortcomings in reliability potentially based on this item, the widespread use of the Net Promoter Score and strong demand to compare website scores using a common metric makes its inclusion justified. Many organizations have adopted the Net Promoter Score as *the* metric to track all activities by, from sales, call center activities, product experiences, and website experiences. In many cases, not including a Net Promoter Score as part of a measurement system is a "deal-breaker."

Research question 4

Research question 4 asked whether the new instrument demonstrates adequate construct and content validity. Studies 1 through 4 showed that the SUPR-Q items exhibited strong convergent validity with the SUS and WAMMI. Study 4 showed that the final version of the SUPR-Q also had strong convergent validity with the Attitude toward the Site (Ast) instrument. The combined score had a high correlation (*r* = .84), and the subfactors also exhibited strong correlations (*r* > .58), suggesting that the SUPR-Q score overall, and to a lesser extent the subfactors, tends to measure a similar construct as other instruments that are used to measure perceptions of the overall quality of the website experience.

There were high correlations between usability and appearance across Studies 1 through 4, suggesting a comingled relationship. This strong correlation was seen with the usability factor of both the SUPR-Q and the SUS. It suggests that participants are rating more attractive websites as more usable. Future analysis should continue to examine the relationship between appearance and usability, similar to Tuch et al. (2012).

There was reasonable evidence for content validity. Items rated by the judges as the easiest to agree to also tended to be the ones that were agreed to more by participants. The average rating from three independent experts in user experience and interface evaluation correlated (rho = .44) with the average responses from the data in Study 4.

Research question 5

Research question 5 asked if users' attitudes toward website quality (as measured by a validated instrument) can be predicted by a calibrated experience checklist used by experts (the CEG).

The average of the 105 items from the initial CEG had a moderate correlation with the SUPR-Q (*r* = .50), suggesting a modest overlap between how judges view a website and how users view the website. The correlation is not strong enough to suggest that the

CEG is a replacement for the SUPR-Q. It does suggest that the CEG and SUPR-Q measure complementary and somewhat overlapping constructs of the quality of the website user experience.

The lower correlation is also likely a function of at least three things. First, the items may be focusing on parts of the website that users are not necessarily encountering during their usage. Second, the items in the CEG may address elements of the website user experience that have an attenuated impact on user attitudes toward the user experience quality. Finally, the variation in participants' prior experience with the site may mask the relationship. That is, users with low experience on websites are likely to be more influenced by website elements (Cavallin, Martin, & Heylighen, 2007). Future analyses can examine how well the refined CEG of 25 items can predict participants' attitudes while accounting for prior experience.

Additionally, both the Winsteps analysis (single facet) and Facets (multiple facets) analysis showed that judges were not fully using the five points of the scale. This is likely due to the wording of many of the items, as some judges may use a "present" or "absent" type of scoring, which would correspond to a 5 for "present" and less than 4 for "absent." After collapsing the response options from five to three, the data fit the model better.

Research question 6

Research question 6 asked whether the effects of the judge experience levels affect the CEG ratings. The Facets analysis revealed that while there was not a difference between the type of Mechanical Turk evaluator (master vs. non-master), there was a significant difference between the experts and evaluators used on Mechanical Turk. The major difference was that the experts had both more experience and more instruction in judging the website and used the entire CEG when evaluating the website. This suggests that if independent evaluators are used, specific training and practice will be needed to bring the scores more in line with more highly trained evaluators.

In practice, checklists are often used by a single evaluator, since it can be difficult to find multiple experienced judges. This analysis suggests that the score from the CEG is dependent on both individual judgment and that judge’s training. When multiple evaluators are used, the Facets analysis suggests that there is some advantage in controlling for the individual judgment of the raters.

Conclusion

Across four studies, more than 4,000 responses to experiences with more than 100 websites were analyzed to generate an eight-item measure of the quality of the website user experience. The questionnaire is called the Standardized User Experience Percentile Rank Questionnaire (SUPR-Q) and contains four factors: usability, trust, appearance, and loyalty.

The factor structure was replicated across the four studies, with data collected both during a usability test and retrospectively. There was evidence of convergent validity with existing questionnaires SUS and WAMMI. The overall average score was shown to have high internal consistency reliability ($\alpha = .86$), while the subscales had lower but generally acceptable levels of reliability ($\alpha = .64$ to $\alpha = .88$). The lower reliability is a consequence of using only two items per factor to keep the total length short—one of the primary goals of this research. Finally, an initial distribution of scores across the websites generated a database to generate percentile ranks and make scores more meaningful to researchers.

To administer the SUPR-Q, users responded to seven of the eight items using a five-point scale (1 = strongly disagree and 5 = strongly agree). For one item ("How likely are you to recommend this website to a friend or colleague?"), users responded to an 11-point scale (0 = not at all likely and 10 = extremely likely). The following are the eight items in the SUPR-Q and their corresponding factor:

This website is easy to use. (usability)

It is easy to navigate within the website. (usability)

The information on this website is trustworthy. (trust)

The information on this website is credible. (trust)

How likely are you to recommend this website to a friend or colleague? (loyalty)

I will likely visit this website in the future. (loyalty)

I find the website to be attractive. (appearance)

The website has a clean and simple presentation. (appearance)

The means and standard deviations derived from Study 3 can be used as the basis for identifying percentile ranks for the overall score and subscale scoring. For example, a website that obtains an overall mean score of 4.1 would be about .6 standard deviations above the mean. This would place it higher than 70 percent of websites in the database, assuming a normal distribution of SUPR-Q scores. Its score can then be expressed as a 70, meaning a percentile rank of 70. While the shapes of the distributions are reasonably normal, additional data may skew the values more or reduce skewness. Future analysis will need to examine the distributions to determine if a log-transformation is needed to maintain a normal distribution. This is essential for computing accurate normed scores.

To complement the SUPR-Q and provide more detailed information on what to fix to improve a website user experience, 48 items, reduced to 37, were identified from an initial pool of 105 items. This checklist can be administered using a three-point scale and is called the Calibrated Evaluator’s Guide (CEG). Evaluators should be trained in judging the website user experience, since a difference was found between experienced and less experienced evaluators and between evaluators more generally. In practice, the 37 items identified in the CEG can be used as an initial screen using multiple judges (with training). For websites needing a more detailed analysis of their strengths and weaknesses, the full 105-item CEG can be administered. Finally, qualified participants who have used the website can answer the SUPR-Q to provide a more comprehensive view of the website user experience. Future research can examine if the further reduced set of items (37 reduced to 18 after identifying similar items) is sufficient for an initial screen. The SUPR-Q and CEG together provide a psychometrically validated picture of the quality of the website user experience. They provide both a score on how well the website is performing relative to a normative dataset and what things need to be addressed to improve the score.

There are existing instruments that are similar to the SUPR-Q, for example the WAMMI (Kirawoski & Cierlik, 1998) and the QUIS (Chin et al., 1988). The SUPR-Q uses common constructs identified in the literature to measure the quality of the user experience: usability (including navigation ease), trust, appearance, and loyalty. There are four reasons why existing measures are inadequate and a new measure is needed. Some existing measures are proprietary instruments (e.g., ForeSee, WAMMI); many do not provide enough coverage of website quality (e.g., only usability with the SUS or loyalty with the NPS); others are too long (e.g., WebQual, WAMMI, SUMI, SUS); and many have insufficiently documented psychometric properties (e.g., CxPI and ForeSee). The CEG complements these shorter customer-facing instruments with a more detailed set of guidelines to be used by trained evaluators, which improves the quality of the website inspection process (Bastien & Scapin, 1995). Using both instruments, data can be collected and compared across disparate products and websites (Whiteside et al., 1988) using two complementary methods and with multiple evaluators, who tend to be more reliable than even a single expert evaluator (Nielsen & Molich, 1990). However, as can be seen from the results, when multiple evaluators are enlisted to evaluate websites, an analysis that incorporates evaluator as a facet is needed.

There are many choices for purchasing consumer retail products or for researching any product or service online. If a user cannot find information, purchase a product easily, or does not trust the information, the user goes elsewhere and may tell friends and colleagues about the poor experience. In practice, to assess and improve the quality of a website user experience, the following process can be followed. First, independent evaluators can be identified to score a website using the 37 items on the CEG. The evaluators should have training in identifying compliance or violation of the items listed on the CEG. The scores of the multiple evaluators will be averaged to create a composite or Rasch analysis using a Facets model. This can then be compared against the scores from the 68 websites in this analysis. A larger database can be maintained to create a more diverse and robust set of scores. Next, qualified participants can be recruited to answer the eight SUPR-Q items after reflecting on their most recent experience with the website. The two scores together will provide a reliable and valid measure of the strengths and weaknesses of the website and what to fix. After changes have been implemented, the CEG and SUPR-Q instruments can be administered again to determine if the website has a quantifiably better user experience.

Limitations and directions for future research

Results of the studies reported herein were limited by the websites evaluated, the experience and biases of the evaluators, and the participants used across the studies. The participants came primarily from the United States, spoke English, and were willing to participate in online studies for compensation. This limited geographical and cultural representativeness may limit the generalizability of the findings to other cultures and geographies.

Future analysis should continue to investigate better items for the CEG. As websites change, new items will likely differentiate websites with excellent user experiences from those with average or poor ones. This will likely render some items obsolete. The work by Armstrong, Green, and Graefe (2016) suggests there can be successful attempts to predict empirical customer attitudes from core design principles as is done with the CEG. This work also lends support for the inclusion of novices in evaluating website user experience quality.

The inclusion of the Net Promoter Score item should also be investigated. If its use wanes in organizations, its inclusion may no longer be warranted, and a replacement that exhibits higher reliability may be necessary. This may be the case especially if additional research finds a weakened link between future growth and likelihood to recommend as found in Keiningham et al. (2007).

Additional analysis should also continue to investigate the relationship between the user experience as measured by the CEG and SUPR-Q. This analysis found only a modest correlation (r = .5) between the two measures. The additional analysis can examine a larger dataset and control for extraneous variables such as prior experience, brand attitude, evaluator judgment, and the effects of the task on the SUPR-Q. A larger set of websites will also allow for analysis by type of website (e.g., industry, non-profit, or for-profit) to understand how items across the factors may change as a function of website type.

Additional items should also be examined for the SUPR-Q, especially when discriminating between websites with high-user experience. A future study can examine additional items using a Rasch analysis to continue to identify items that create a "yardstick" of items from the low to high end of user experience quality.

Finally, the databases for the SUPR-Q and CEG should be built upon a larger and more diverse set of websites. This should also focus on non-ecommerce websites such as government and nonprofit sites—websites where there's likely a lack of effort on measuring and improving the quality of the user experience. Additional factors such as translation and globalization (tests across countries) should be investigated to understand how language and culture may impact the reliability and validity of these measures.

REFERENCES

Abdul-Muhmin, A. (2010). Repeat purchase intentions in online shopping: The role of satisfaction, attitude, and online retailers' performance. *Journal of International Consumer Marketing*, *23*(1), 5–20. doi:10.1080/08961530.2011.524571.

Aladwani, A. M., & Palvia, P. C. (2002). Developing and validating an instrument for measuring user perceived web quality. *Information Management*, *39*(6), 467–476. doi:10.1016/S0378-7206(01)00113-6.

Angriawan, A., & Thakur, R. (2008). A parsimonious model of the antecedents and consequence of online trust: An uncertainty perspective. *Journal of Internet Commerce*, *7*(1), 74–94. doi:10.1080/15332860802004337.

Apple Computer. (1987). *Human interface guidelines: The Apple desktop interface*. Reading, MA: Addison-Wesley.

Armstrong, S., Du, R., Green, K., & Graefe, A. (2016). Predictive validity of evidence-based persuasion principles: An application of the index method. *European Journal of Marketing, 50*(1-2), 276.

Bargas-Avila, J. A., Lötscher, J., Orsini, S., & Opwis, K. (2009). Intranet satisfaction questionnaire: Development and validation of a questionnaire to measure user satisfaction with the intranet. *Computers in Human Behavior*, *25*(6), 1241–50. doi:10.1016/j.chb.2009.05.014.

Bevan, N. (2009). Extending quality in use to provide a framework for usability measurement. In M. Kurosu (Ed.), *Human Centered Design* (pp. 13–22). Heidelberg, Germany: Springer Berlin.

Bobko, P. (2001). *Correlation and regression: Applications for industrial organizational psychology and management*. New York, NY: Sage.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge.

Boostrom, R., Balasubramanian, S. K., & Summey, J. H. (2013). Plenty of attitude: Evaluating measures of attitude toward the site. *Journal of Research in Interactive Marketing, 7*(3), 201–15. doi:10.1108/JRIM-02-2013-0012.

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the system usability scale: A test of alternative measurement models. *Cognitive Processes*, *10*(3), 193–197. doi:10.1007/s10339-009-0268-9.

Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189–194). London, UK: Taylor & Francis.

Bruner, G. C., II, & Kumar, A. (2002). Similarity analysis of three attitude-toward-the-website scales. *Quarterly Journal of Electronic Commerce*, *3*(2), 163–172.

Cavallin, H., Martin, W. M., & Heylighen, A. (2007). How relative absolute can be: SUMI and the impact of the nature of the task in measuring perceived software usability. *AI & Society*, *22*(2), 227–235. doi:10.1007/s00146-007-0127-0.

Chen, Q., & Wells, W. D. (1999). Attitude toward the site. *Journal of Advertising Research, 39*(5), 27–37.

Chen, Q., Clifford, S. J., & Wells, W. D. (2002). Attitude toward the site II: New information. *Journal of Advertising Research, 42*(2), 33–45. doi:10.2501/JAR-42-2-33-45.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human–computer interface. In E. Soloway, D. Frye, & S. Sheppard (Eds.), *Proceedings of the ACM CHI 88 Human Factors in Computing Systems Conference* (pp. 213–18). Washington, DC: ACM. doi:10.1145/57167.57203

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–339. doi:10.2307/249008.

Dumas, J., & Redish, J. C. (1999). *A practical guide to usability testing*. Portland, OR: Intellect.

Enright, A. (2014). U.S. online retail sales will grow 57% by 2018; projected growth.

Retrieved from http://www.internetretailer.com/2014/05/12/us-online-retail-sales-will-grow-57-2018.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers, 22*(5), 323–27. doi:10.1016/j.intcom.2010.04.004.

Flores, L. (2004, February). 10 facts about the value of brand websites. *Admap*, 26–28.

ForeSee. (2015). Industry and competitive benchmarks. Retrieved from http://www.foresee.com/services/benchmark-and-compare/.

George, J. F. (2002). Influences on the intent to make internet purchases. *Internet Research*, *12*(2), 165–80. doi:10.1108/10662240210422521.

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human Computer Interaction*. *13*(4), 481–99. doi:10.1207/S15327590IJHC1304_07.

Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology, 25*(2), 91–97. doi:10.1080/01449290500330331.

Hollingsed, T., & Novick, D. (2007). Usability inspection methods after 15 years of research and practice. In *Proceedings of the 25th Annual ACM International Conference on Design of Communication*. New York, NY: ACM. doi:10.1145/331490/331493.

Hong, S., & Kim, J. (2004). Architectural criteria for website evaluation - conceptual framework and empirical validation. *Behaviour & Information Technology, 23*(5), 337-357. doi:10.1080/01449290410001712753

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*(2), 79–102. doi:10.1016/j.ijhcs.2005.06.002.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453. doi:10.1037//1082-989X.3.4.424.

International Organization for Standardization. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs), part 11, guidance on usability* (ISO 9241-11:1998E). Geneva, Switzerland: ISO.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques. In S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 119–124). New York: ACM. doi:10.1145/108844.108862.

John, B. E., & Marks, S. J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, *16*(4/5), 188–202. doi:10.1080/014492997119789.

Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In P. Bauersfeld, J. Bennett, & G. Lynch (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 397–404). New York: ACM. doi:10.1145/142750.142873.

Keeney, R. L. (1999). The value of internet commerce to the customers. *Management Science*, *45*(4), 533–542. doi:10.1287/mnsc.45.4.533.

Keiningham, T. L., Cooil, B., Andreassen, T. W., & Aksoy, L. (2007). A longitudinal examination of net promoter and firm revenue growth. *Journal of Marketing, 71*(3), 39-51. doi:10.1509/jmkg.71.3.39

Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and usage. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 169–78). London, UK: Taylor & Francis.

Kirakowski, J., & Cierlik, B. (1998). Measuring the usability of websites. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 424–428). Santa Monica, CA: HFES.

Kline, R. B. (2011). *Principles and practices of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Korgaonkar, P., & Wolin, L. D. (2002). Web usage, advertising, and shopping: Relationship patterns. *Internet Research, 12*(2), 191–204. doi:10.1108/10662240210422549.

Krug, S. (2014). *Don’t make me think, revisited: A common sense approach to web usability* (3rd ed.). Berkeley, CA: New Riders.

Lascu, D.-N., & Clow, K. E. (2008). Web site interaction satisfaction: Scale development considerations. *Journal of Internet Commerce, 7*(3), 359–78. doi:10.1080/15332860802250476.

Law, E. L.-C., & Hvannberg, E. T. (2004). Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In *Proceedings of the Third*

Nordic Conference on Human-Computer Interaction (pp. 241–50). New York, NY: ACM. doi:10.1145/1028014.1028051.

Lewis, C., Polson, P., Wharton, C., & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People*. New York, NY: ACM. doi:10.1145/97243.97279.

Lewis, J., Utesch, B., & Maher, D. (2013). UMUX-LITE: When there's no time for the SUS. In *Proceedings of the Conference in Human Factors in Computing Systems* (pp. 2099–2102). New York, NY: ACM. doi:10.1145/2470654.2481287.

Lewis, J. R. (1992). Psychometric evaluation of the Post-Study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1259–63). Santa Monica, CA: Human Factors Society. doi:10.1177/154193129203601617.

Lewis, J. R. (2012a). Predicting Net Promoter scores from System Usability Scale scores. Available at www.measuringu.com/blog/nps-sus.php (accessed April 4, 2014).

Lewis, J. R. (2012b). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (4th ed., pp. 1267–1312). New York, NY: John Wiley.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2015). *Facets computer program for many-facet Rasch measurement, version 3.71.4.* Beaverton, OR: Winsteps.com

Lohse, G., & Spiller, P. (1999). Internet retail store design: How the user interface influences traffic and sales. *Journal of Computer-Mediated Communication, 5*(2). doi:10.1111/j.1083-6101.1999.tb00339.x.

Loiacono, E. T., Watson, R. T., & Goodhue, D. L. (2002). WebQual™: A measure of website quality. *Marketing Theory and Applications*, *13(3)*, 432–38.

Lo Storto, C. (2013). Evaluating ecommerce websites cognitive efficiency: An integrative framework based on data envelopment analysis. *Applied Ergonomics, 44*(6), 1004. doi:10.1016/j.apergo.2013.03.031

Microsoft Corporation. (1995). *The Windows interface: Guidelines for software design*. Redmond, WA: Microsoft Press.

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Academic Press.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 152–158). Boston, MA: ACM. doi:10.1145/191666.191729.

Nielsen, J. (1995). 10 usability heuristics for user interface design. Retrieved from http://www.nngroup.com/articles/ten-usability-heuristics/.

Nielsen, J. (2001). 113 design guidelines for homepage usability. Retrieved from http://www.nngroup.com/articles/113-design-guidelines-homepage-usability/.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People* (pp. 249–256). New York, NY: ACM. doi:10.1145/97243.97281.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Pavlou, P. A., & Fygenson, M. (2006). Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior. *MIS Quarterly, 30*(1), 115–143. doi:10.2307/25148720.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded ed.). Chicago: University of Chicago Press.

Reichheld, F. (2006). *The ultimate question: Driving good profits and true growth*. Boston, MA: Harvard Business School Press.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*, 46–54.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (2nd ed.). Boston, MA: Wiley.

Safar, J. A., & Turner, C. W. (2005). Validation of a two factor structure of system trust. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, *49*(3), 497–501. doi:10.1177/154193120504900360.

Sauro, J. (2010a). Does better usability increase customer loyalty? Retrieved from www.measuringu.com/usability-loyalty.php.

Sauro, J. (2010b). How many users do people actually test? Retrieved from http://www.measuringu.com/blog/actual-users.php.

Sauro, J. (2011). *A practical guide to the System Usability Scale (SUS): Background, benchmarks & best practices*. Denver, CO: Measuring Usability LLC.

Sauro, J. (2012). How effective are heuristic evaluations? Retrieved from http://www.measuringu.com/blog/effective-he.php.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of ACM CHI 2009 Conference on Human Factors in Computing Systems* (pp. 1609–18). Boston, MA: ACM. doi:10.1145/1518701.1518947.

Sauro, J., & Lewis J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of ACM CHI 2011 Conference on Human Factors in Computing Systems* (pp. 2215–2223). Vancouver, BC, Canada: ACM. doi:10.1145/1978942.1979266.

Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Waltham, MA: Morgan Kaufmann.

Scapin, D. L., & Bastien, J. M. C. (1997). Ergonomic criteria for evaluating the ergonomic quality of interactive systems. *Behaviour & Information Technology*, *16*(4), 220-231. doi:10.1080/014492997119806

Smith, S., & Mosier, J. (1986). *Guidelines for designing user interface software*. Report MTR.10090. Bedford, MA: The MITRE Corp.

Stanford University. (2004). *Stanford guidelines for web credibility*. Retrieved from https://credibility.stanford.edu/guidelines/.

Statistica. (2012). Why do online shoppers leave without paying? Retrieved from http://www.statista.com/statistics/232285/reasons-for-online-shopping-cart-abandonment/.

Stewart, T. (2015). User experience. *Behaviour & Information Technology*, *34*(10), 949–951. doi:10.1080/0144929X.2015.1077578

Stone, J. (2015, July 3). Top 10 US retailers: Amazon joins ranks of Walmart, Kroger for first time ever. Retrieved from http://www.ibtimes.com/top-10-us-retailers-amazon-joins-ranks-walmart-kroger-first-time-ever-1618774.

Suh, B., & Han, I. (2003). The impact of customer trust and perception of security control on the acceptance of electronic commerce. *International Journal of Electronic Commerce*, *7*(3), 135–61. doi:10.1080/10864415.2003.11044270.

Supphellen, M., & Nysveen, H. (2001). Drivers of intention to revisit the websites of

well-known companies. *International Journal of Market Research*, *43*(3),

3

41–52.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Allyn and Bacon.

Travis, D. (2009). Web usability guidelines. Retrieved from http://www.userfocus.co.uk/resources/guidelines.html.

Trefis Team. (2014, June 9). Wal-Mart expects 30% rise in e-commerce revenues this year. Retrieved from http://www.forbes.com/sites/greatspeculations/2014/06/09/wal-mart-expects-30-rise-in-e-commerce-revenues-this-year/.

Tuch, A. N., Roth, S. P., Hornbæk, K., Opwis, K., & Bargas-Avila, J. A. (2012). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior*, *28*(5), 1596–1607. doi:10.1016/j.chb.2012.03.024.

Tullis, T., & Albert, B. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (2nd ed.). Boston; Amsterdam: Elsevier/Morgan Kaufmann.

Tullis, T. S., & Stetson, J. N. (2004). A comparison of questionnaires for assessing website usability. Paper presented at the *Usability Professionals Association Annual Conference* (pp. 1–12). Minneapolis, MN: UPA.

United States Census Bureau. (2015, May 15). US Census Bureau News. Retrieved from https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf.

United States Department of Health and Human Services. (2006). *The research-based web design & usability guidelines, enlarged/expanded edition*. Washington: U.S. Government Printing Office.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. doi:10.2307/30036540.

Wang, J., & Senecal, S. (2007). Measuring perceived website usability. *Journal of Internet Commerce, 6*(4), 97–112. doi:10.1080/15332860802086318.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability engineering: Our experience and evolution. In M. Helander (Ed.), *Handbook of Human Computer Interaction* (pp. 791–818). New York: North Holland.

Wolfinbarger, M., & Gilly, M. C. (2002). .comQ: Dimensionalizing, measuring and predicting quality of the e-tail experience. *Marketing Science Institute Report* No. 02-100. Cambridge, MA: Marketing Science Institute.

Zeithaml, V. A., Parasuraman, A., & Malhotra, A. (2000). A conceptual framework for understanding e-service quality: Implications for future research and managerial practice. *MSI Working Paper Series*, Report No. 00-115. Cambridge, MA: Marketing Science Institute.

APPENDIX A: Nielsen’s 10 Heuristics

Nielsen’s 10 Heuristics (Nielsen, 1995)

1. Visibility of system status

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

2. Match between system and the real world

The system should speak the users’ language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

3. User control and freedom

Users often choose system functions by mistake and will need a clearly marked “emergency exit” to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

4. Consistency and standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

5. Error prevention

Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

6. Recognition rather than recall

Minimize the user’s memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

7. Flexibility and efficiency of use

Accelerators—unseen by the novice user—may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

8. Aesthetic and minimalist design

Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

9. Help users recognize, diagnose, and recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

10. Help and documentation

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user’s task, list concrete steps to be carried out, and not be too large.

APPENDIX B: Tasks Used in Study 4

Real Estate Task: Search for a single-family home in Denver, Colorado. The home should be between $325,000 and $400,000 and have three bedrooms and two bathrooms. Of the houses that fit the criteria, please select a home that matches your needs/wants the MOST.

Write down the address of the home, as you will be asked for it later.

Video Streaming Task: Let’s imagine you’re at home and are looking for an action movie to watch. Using [website], find an **action movie** you’d be interested in watching and make sure it has between a **four- and five-star rating**. Write down the name of the movie, as you will be asked for that information later.

APPENDIX C: Tasks Used in Study 5

Website	Task
Chipotle	Imagine you are interested in ordering food online from Chipotle.com. Find the subtotal for a Chicken Burrito Bowl and one Bottled Water and write down or copy that price.
Chipotle	Suppose you want to know the nutrition information for some of Chipotle's food. Find the total calories for chips and guacamole and copy or write it down.
eBay	Imagine that you are interested in ordering an eBay gift card as a small gift. Find and copy or write down the lowest amount offered for an eBay gift card.
eBay	Imagine that you have an item you want to sell on eBay and you want to know if they charge any fees for selling on their website. Find and copy or write down the fee charged by eBay on sales from a standard eBay seller account.
Microcenter	Imagine that you are researching a certain desktop computer that you are interested in purchasing from Microcenter. Find and copy or write down the price for a Lenovo H30 Desktop Computer.
Microcenter	Suppose your laptop has a broken screen and you are interested in getting it fixed. Find and copy or write down the starting price for a Laptop Screen Repair.
Enterprise	Imagine you are going to be in Los Angeles from July 3rd to July 5th. You will be arriving and leaving from the Los Angeles International Airport. Find and copy or write down the total price of a rental vehicle for that time period if you are renting an Economy Class vehicle. (For the purposes of this question, assume you are 25 or older.)
Enterprise	Suppose that you want to save some of Enterprise's emergency information for while you are renting one of their cars. Find and copy or write down the Enterprise phone number for Roadside Assistance.
ULWorkplace	Suppose you are interested in taking some of the PureSafety courses offered by ULWorkplace. Find and copy or write down the starting price for PureSafety courses for individuals.
ULWorkplace	Suppose a colleague recommended a ULWorkplace webinar that featured an interesting presenter. Find and copy or write down the name of the featured presenter for the live webinar that occurred on March 4, 2014, titled "The Economics of Ergonomics."
BicycleDoctorUSA.com	Suppose that you are researching a new bicycle and are interested in a specific bike on BicycleDoctorUSA.com. Find and copy or write down the starting price of a 2015 Kestrel RT-1000 Carbon Fiber road bike.
BicycleDoctorUSA.com	Imagine you are ordering a small item from BicycleDoctorUSA.com and want to know how much shipping and handling will be. Find and copy or write down the standard shipping and handling prices for small orders.
Craigslist	Suppose that you are interested in buying a used laptop and are researching MacBook Pro laptops in the Denver area. Find and copy or write down the range of prices you might expect to pay for a non-retina MacBook Pro laptop.

Craigslist	Imagine that you are researching job positions in the Denver area. How many job listings were posted for Legal/Paralegal positions in the Denver area on May 28th, 2015?
Etsy	Suppose you know a friend likes a certain seller on Etsy.com and you want to buy her a ceramic pitcher as a gift. Find and write down the price of a pitcher made by PigeonToeCeramics.
Etsy	Imagine that you are an artist and you want to list your work for sale on Etsy. Find and copy or write down the listing fee for posting items on the website.
Newsweek	Imagine you looking for the featured article of a particular issue of *Newsweek*. Find and write down or copy the title of the cover article for *Newsweek's* May 15, 2015, issue.
Newsweek	Imagine you remember a good article that you read on newsweek.com, and you want to find out who the author was. Find and write down or copy who wrote the article titled "Everything You Need to Know About the Belmont Stakes."
Wired	Imagine you remember a good article that you read on wired.com, and you want to find out who the author was. Find and write down or copy who wrote the article titled "Microsoft Says Windows 10 Will Arrive on July 29."
Wired	Suppose you are interested in subscribing to Wired Magazine and want to know how much it will cost. Find and write down or copy the current price for a subscription to *Wired* magazine for six months.
Macy's	Suppose you are comparing jackets to order as a gift and want to know the price of a certain fleece jacket. Find and write down or copy the original price (NOT the sale price) for a Champion Duofold Quarter-Zip Fleece Jacket.
Macy's	Imagine you are interested in the range of watches that Macy's carries. Find and write down or copy the price of the most expensive women's watch available from macys.com.
Home Depot	Suppose you are looking to buy a specific refrigerator and want to know how much it would cost from the Home Depot. Find and write down or copy the price for a Frigidaire 15 cu. ft. Top Freezer Refrigerator in Stainless Steel.
Home Depot	Suppose that you need to call a specific Home Depot location to ask some questions. Find and write down or copy the phone number for the Home Depot store on 1600 29th Street in Boulder, Colorado.
AirMac	Imagine that you are interested in a certain brand of air compressors and want to know if they are carried by AirMac. Find and write down or copy how many Champion brand reciprocating air compressors are listed on airmac.com.
AirMac	Imagine you want to know what brands of air compressors AirMac is able to service. Find and write down or copy how many brands AirMac services through their Air Compressor Field Service and Repair program.
Harbor Freight	Imagine that you want to order some new garden tools from Harbor Freight. Find and write down or copy the original price (NOT the sale price) for a four-piece garden tool set.
Harbor Freight	Suppose that you are browsing Harbor Freight's current sales. Find and write down or copy the current sale price for the first item listed on the monthly ad.
Adobe	Suppose you want to buy a set of software from Adobe. Find and write down or copy the price to buy Photoshop Elements and Premiere Elements together.

Adobe	Imagine that you are interested in purchasing some photography software from Adobe. Find and write down or copy which products are included in the Creative Cloud Photography Plan.
Dell	Imagine that you are interested in buying a new monitor for your computer and are looking into purchasing a touch-screen monitor. Find and write down or copy the price of a Dell 23 Touch Monitor.
Dell	Suppose that you are interested in purchasing a certain computer and have some specific hardware requirements. Find and write down or copy the price for an Inspiron 20 3000 Series all-in-one computer, with a dual core processor and a touch screen.

APPENDIX D: CEG Checklist

* Marked items are retained in final Checklist from Winsteps.

+ Marked items are retained in final Checklist from Facets.

Navigation

1. Navigation options are ordered in a logical or task-oriented manner.*+
2. Navigational paths are direct and reasonable in length.*
3. Users are aware of where they are in the website.*
4. The site is organized in an understandable and navigable manner.*
5. It is easy to leave or deviate from a navigational path, but it is clear when a user is doing so.*+
6. Button and link labels effectively indicate where the user will be taken.*+
7. Navigational options are visible and obvious.*+
8. There is no unnecessary scrolling.+
9. Links do not take users to external sites without alerting the user.
10. The pace of the navigation does not rush the user or hold them back from accessing content.*+
11. The website's navigation is usable for the typical user.*+

Information

12. Content is concise and relevant.*+
13. Content is presented in the most appropriate form.*
14. The information is recent and up to date.+
15. Content is laid out in the most logical manner.*
16. Pages are quick to scan and evaluate.*
17. Content is easy to find and get to.*+
18. Pages are free from distractions.*
19. It is easy to locate and identify desired information.*+
20. It is easy to save, print, or download information as appropriate.+

21. Content formatting is effective and visually appealing.+

22. The website’s content is usable for the typical user.*+

Search

23. The default search is intuitive to use.+

24. The search results page shows the user what was searched for and it is easy to edit and re-search.

25. Search results are clearly presented and sorted according to relevance by default.

26. Search results can be filtered and sorted by the user, and the results per page can be customized.

27. The search handles errors, misspellings, and blank searches effectively.

28. The search includes alternate spellings, plurals, and related terms as applicable.

29. There is a more advanced search, and it is easily accessed and intuitive to use.

30. The searches cover an appropriate portion of the website as relevant to the user’s needs.+

31. Search results can be viewed in multiple ways as applicable, e.g., list, thumbnail, etc.

32. The search bar employs predictive text when useful.

33. The website’s search function is usable for the typical user.+

Product Pages

34. Products are categorized effectively and all categories are visible and accessible.

35. Products are sorted and ordered within categories according to the typical user’s needs.

36. Physical products, services, and digital products are all clearly differentiated.

37. Products can be viewed in multiple ways, e.g., lists, thumbnails, etc.

38. Users can sort and filter products based on useful parameters.

39. The basic product information is presented in the category views.

40. All necessary product information is presented in the individual product pages.

41. Product options are clearly available, e.g., color, quantity, size, etc.

42. It is easy to add products to the cart, and it is clearly indicated when this has been done.

43. Products can be easily compared, both as default configurations and user-customized configurations.

44. Reviews, ratings, and other user feedback is available, as applicable.

45. Featured items and deals are visible and effectively formatted.

46. Advanced product customizers are intuitive and comprehensive, e.g., electronics/computers

47. It is clear to users where they are in the website and how they have previously sorted or filtered products.

48. The website's product pages are usable for the typical user.

Purchasing and Billing

49. The critical path is clear to the user, and it is comfortably paced.

50. Users are always made aware of where they are in the transaction process.

51. There is a cart/basket page that summarizes its contents and prices.

52. It is clear to users when they are submitting a purchase and everything that is included in that purchase.

53. Digital receipts, confirmation pages, and product tracking options are presented to the user.

54. Transactions can be saved for later and are not lost if a user leaves the process before completion.

55. Fields for entering customer information are formatted appropriately and indicate what information is required.

56. Forms only require necessary information and are validated at the appropriate time.

57. All costs and fees are clearly presented to the user before initiating a purchase, e.g., totals, shipping and handling, etc.

58. Actions can be undone and errors can be fixed.

59. Products in the cart can be added, removed, or edited.

60. Payment options and security are clearly communicated to the user.

61. Registration is only required when necessary.

62. Additional options (newsletters, email subscriptions, coupon codes) are clear but not disruptive.

63. Cross-selling is used appropriately; not intrusive or easily confused with cart contents.

64. The website's purchasing and payment functions are usable for the typical user.

Forms & Data Entry

65. Fields clearly indicate what information is being requested.

66. Tips or hints are provided if users do not understand what information is required.

67. Fields indicate the format that information must be entered in.

68. Fields automatically format information when it is entered, as applicable.

69. Entry fields are the right size for the information being entered.

70. It is clear what information is required and what information is optional.

71. Forms indicate if external information is required and where to find that information.

72. Data is entered in the most appropriate format (e.g., drop-down menus vs. text entry).

73. Forms automatically complete information based on previous entries (e.g., ZIP code or state).

74. Forms fields are validated at the proper time.

75. Error messages appear at the proper time (e.g., after the user is done entering information in a field).

76. Error messages clearly indicate what went wrong and help the user resolve the error.

77. The site does not ask for sensitive personal information unless it is absolutely necessary.

78. Password fields are obscured or hidden in usable and secure methods.

79. Please rate how usable you feel the website's forms are for the typical user.

Help & Information

80. The website's privacy policy is easily accessible and easy to understand.

81. The FAQ or online help section is clear and helpful for user's needs.

82. It is easy to contact the corporation or organization behind the website for help with the company’s services or assistance with the site.

83. There is a variety of solutions to problems that the users may encounter, e.g., FAQ, online help, tutorials, contact info.

84. All necessary corporate and website information is accessible.

85. Any necessary legal information is accessible and organized logically.

86. Relevant site settings are customizable, e.g., language, location, local store.

87. About Pages contain relevant and useful information and make the website more credible and trustworthy (as applicable).

88. Error messages are worded effectively and direct the user to the appropriate solution.

89. If applicable, bios are provided with relevant contact information/links to outside profiles, e.g., faculty, employees, press.

90. The website’s help and informational components are usable for the typical user.

Overall Elements

91. The website can be utilized by users with varying levels of experience.+

92. The website is functioning correctly, e.g., there are no broken links, missing pages, etc.

93. The website contains a logical selection of relevant pages, e.g., home page, about, products.+

94. The website does not overuse advertisements and popups.

95. The website fully utilizes its resources, e.g., screen space, navigation, etc.+

96. The site does not utilize unnecessary novel devices such as interactive banners or flash-based pages.+

97. Readability throughout the site is clear.

98. The text and background are contrasting colors.

99. Text is in colors and fonts that are readable and appropriate for the site.+

100. The site layout is balanced and symmetrical.+

101. The visual design is appealing to the typical user.+

102. The overall design is relevant to the website’s purpose.

103. The design is not overly flashy or distracting.

104. The visual design is original and related to the website’s other branding.+

105. The website’s navigation is usable for the typical user.*+

APPENDIX E : Comparison of CEG to Usability.gov Guidelines

CEG	Usability.Gov	p.	Section
	Information		
Content is concise and relevant.	Display Only Necessary Information	176	16:07
Content is presented in the most appropriate form.	Organize Information Clearly	170	16:01
	Group Related Elements	173	16:04
	Format Information for Multiple Audiences	177	16:08
The information is recent and up to date.	Ensure that Necessary Information is Displayed	172	16:03
Content is laid out in the most logical manner.	Facilitate Scanning	171	16:02
Pages are quick to scan and evaluate.	Design Quantitative Content for Quick Understanding	175	16:06
	Use Color for Grouping	178	16:09
	Navigation		
Navigational paths are direct and reasonable in length.	Minimize the Number of Clicks or Pages	174	16:05
Users are aware of where they are in the website.	Provide Feedback on User's Location	62	7:04
	Breadcrumb Navigation	70	7:12
It is easy to leave or deviate from a navigational path, but it is clear when a user is doing so.	Use Site Maps	68	7:10
The site is organized in an understandable and navigable manner.	Use Descriptive Tab Labels	64	7:06
Button and link labels effectively indicate where the user will be taken.	Use 'Glosses' to Assist Navigation	69	7:11
Navigation options are ordered in a logical or task-oriented manner.	Use Appropriate Menu Types	67	7:09
Navigational options are visible and obvious.	Provide Navigational Options	59	7:01
	Differentiate and Group Navigation Elements	60	7:02
	Present Tabs Effectively	65	7:07
	Place Primary Navigation Menus in the Left Panel	63	7:05
There is no unnecessary scrolling.	Eliminate Horizontal Scrolling	72	8:01
	Facilitate Rapid Scrolling While Reading	73	8:02
	Use Scrolling Pages for Reading	74	8:03

	Comprehension		
	Use Paging Rather Than Scrolling	74	8:04
	Scroll Fewer Screenfuls	75	8:05
Search			
The default search is intuitive to use.	Provide Search Templates	187	17:09
	Include Hints to Improve Search Performance	186	17:08
The search bar employs predictive text when useful.	Design Search Around Users' Terms	183	17:05
The search results page shows the user what was searched for and it is easy to edit and re-search.	Ensure Usable Search Results	180	17:01
Search results are clearly presented and sorted according to relevance by default.			
Search results can be filtered and sorted by the user, and the results per page can be customized.			
Search results can be viewed in multiple ways as applicable, e.g., list, thumbnail, etc.			
The search handles errors, misspellings, and blank searches effectively.	Make Upper- and Lowercase Search Terms Equivalent	181	17:03
The search includes alternate spellings, plurals, and related terms as applicable.			
There is a more advanced search, and it is easily accessed and intuitive to use.	Notify Users when Multiple Search Options Exist	185	17:07
	Allow Simple Searches	184	17:06
The searches cover an appropriate portion of the website as relevant to the user's needs.	Design Search Engines to Search the Entire Site	181	17:02
Forms			
Fields clearly indicate what information is being requested.	Label Data Entry Fields Consistently	123	13:03
	Label Pushbuttons Clearly	122	13:02
	Label Data Entry Fields Clearly	124	13:05
	Put Labels Close to Data Entry Fields	126	13:07
Tips or hints are provided if users do not understand what information is required.	Do Not Make User-Entered Codes Case Sensitive	123	13:04
Fields indicate the format that information must be entered in.	Label Units of Measurement	135	13:16
	Display Default Values	137	13:18
Entry fields are the right size	Distinguish Required and Optional Data	121	13:01

for the information being entered.	Entry Fields		
It is clear what information is required and what information is optional.			
Data is entered in the most appropriate format (e.g., drop-down menus vs. text entry)	Use Radio Buttons for Mutually Exclusive Selections	128	13:09
	Use Familiar Widgets	129	13:10
	Partition Long Data Items	131	13:12
	Use a Single Data Entry Method	132	13:13
	Prioritize Pushbuttons	133	13:14
	Use Open Lists to Select One from Many	139	13:21
	Use Data Entry Fields to Speed Performance	140	13:22
	Use a Minimum of Two Radio Buttons	140	13:23
	Use Check Boxes to Enable Multiple Selections	134	13:15
	Do Not Limit Viewable List Box Options	136	13:17
	Minimize Use of the Shift Key	141	13:25
Forms automatically complete information based on previous entries (e.g., ZIP code or state)	Minimize User Data Entry	125	13:06
Fields automatically format information when it is entered, as applicable.	Provide Auto-Tabbing Functionality	141	13:24
	Place Cursor in First Data Entry Field	138	13:19
Forms fields are validated at the proper time.	Anticipate Typical User Errors	130	13:11
Error messages appear at the proper time (e.g., after the user is done entering information in a field)	Ensure that Double-Clicking Will Not Cause Problems	138	13:20
Error messages clearly indicate what went wrong and help the user resolve the error.			
Password fields are obscured or hidden in usable and secure methods.	Allow Users to See Their Entered Data	127	13:08
Overall			
The website does not overuse advertisements and popups.	Avoid Cluttered Displays	45	6:01
The site does not utilize unnecessary novel devices such as interactive banners or flash based pages.	Optimize Display Density	50	6:06
The design is not overly flashy or distracting.	Use Moderate White Space	55	6:11
The site layout is balanced and symmetrical.	Place Important Items Consistently	46	6:02
	Place Important Items at Top Center	47	6:03
	Align Items on a Page	51	6:07
	Structure for Easy Comparison	48	6:04
	Establish Level of Importance	49	6:05

The website fully utilizes its resources, e.g., screen space, navigation, etc.	Use Fluid Layouts	52	6:08
	Avoid Scroll Stoppers	53	6:09
	Set Appropriate Page Lengths	54	6:10
	Choose Appropriate Line Lengths	56	6:12
Readability throughout the site is clear.	Use Black Text on Plain, High-Contrast Backgrounds	101	11:01
The text and background are contrasting colors.	Color-Coding and Instructions	108	11:09
Text is in colors and fonts that are readable and appropriate for the site.	Ensure Visual Consistency	103	11:04
The visual design is appealing to the typical user.	Use Bold Text Sparingly	104	11:05
The overall design is relevant to the website’s purpose.	Use Attention-Attracting Features when Appropriate	105	11:06
The visual design is original and related to the website’s other branding.	Format Common Items Consistently	102	11:02
The website can be utilized by users with varying levels of experience.	Use Familiar Fonts	106	11:07
The website is functioning correctly, e.g., there are no broken links, missing pages, etc.	Use at Least 12-Point Font	107	11:08
	Highlighting Information	110	11:11
	Use Frames when Functions Must Remain Accessible	57	6:13
	Use Mixed-Case for Prose Text	102	11:03
The website contains a logical selection of relevant pages, e.g., home page, about, products.	Emphasize Importance	109	11:10

APPENDIX F: Websites and Number of Evaluators Using CEG

Website	# of Evaluators
FootLocker	12
J.Crew	12
JCPenney	12
Walmart	12
Amway	11
Coach	11
OneKingsLane	10
Priceline	10
Nordstrom	9
RestorationHardware	9
BestBuy	8
Dell	8
EddieBauer	8
Expedia	8
NET-A-PORTER	8
SHOP.COM	8
Apple	7
EsteeLauder	7
Nike	7
Target	7
TigerDirect.com	7
Blair	6
Lenovo	6
ABMC	5
Amazon	5
Gilt	5
HP	5
Orbitz	5
OrientalTradingCompany	5
Abercrombie&Fitch	3

AdvanceAutoParts	3
DOJOJP	3
DOTRITA	3
EdibleArrangements	3
Express	3
FAA	3
FDA	3
GAO	3
HRSA	3
KeurigGreenMountain	3
L.L.Bean	3
NASA	3
QVC	3
REI	3
SaksFifthAvenue	3
SEC	3
SSA	3
STATEDEPT	3
TheChildren'sPlace	3
VA	3
Wayfair	3
Williams-Sonoma	3
Adobe	2
AirMac	2
Grainger	2
HarborFreight	2
HomeDepot	2
Macy's	2
Newsweek	2
Wired	2
BicycleDoctorUSA	1
Chipotle	1
Craigslist	1
eBay	1
Enterprise	1

Etsy	1
Microcenter	1
ULWorkplace	1

APPENDIX G: Facets Output

Table G-1

Judge Measurement Report (Arranged by MN)

tal	Total	Obsvd	Fair(M		Mod el	Infi	**T**	Outf	it	Esti m	Corre	lation	Exac t	Agree.	
Score	**Count**	**Average**	**Averag**	**Measure**	**S.E.**	**MnSq**	**ZStd**	**MnSq**	**ZStd**	**Discr**	**PtMea**	**PtExp**	**Obs %**	**Exp %**	**Judge #**
30	11	2.73	3	-6.18	0.65	0.7	-0.4	0.64	-0.5	1.23	0.57	0.25	45.5	43.1	119
28	11	2.55	2.99	-5.5	0.54	1.57	1.3	2.1	2.1	0.25	-0.24	0.3	45.5	45.1	121
57	23	2.48	2.99	-5.11	0.35	1.2	0.8	1.31	1.1	0.73	0.12	0.24	20	22.3	214
28	10	2.8	2.99	-4.58	0.77	1.88	1.3	1.26	0.5	0.85	0.49	0.26	28.6	39.5	62
11	5	2.2	2.98	-4.34	0.68	0.24	-2	0.26	-2	2.48	0.8	0.21	20	40.8	223
91	37	2.46	2.98	-4.27	0.28	1.05	0.3	1.11	0.5	0.87	0.03	0.25	32.4	40.1	165
29	10	2.9	2.98	-4.17	1.04	1.09	0.4	1.52	0.7	0.88	-0.25	0.2	20	28.9	55
28	14	2	2.98	-4.09	0.4	1.8	2.2	1.78	2.2	-0.49	0.31	0.29	36.4	33.4	47
27	10	2.7	2.98	-4.08	0.65	0.59	-0.7	0.6	-0.6	1.33	0.73	0.29	42.9	40.5	59
27	10	2.7	2.98	-4.08	0.65	1.32	0.7	1.03	0.2	0.98	0.66	0.29	35.7	40.5	208
87	37	2.35	2.98	-3.98	0.26	1.52	2.3	1.46	2.1	0.24	0.18	0.26	33.8	41	153
60	22	2.73	2.97	-3.86	0.45	1.41	1.1	1.15	0.4	0.95	0.63	0.21	48.8	40.1	201
67	23	2.91	2.97	-3.85	0.73	1.03	0.2	1.17	0.4	0.95	-0.1	0.14	50.9	48.8	128
91	32	2.84	2.97	-3.74	0.48	1.12	0.4	0.71	-0.5	1.1	0.67	0.19	40	37	149
153	56	2.73	2.96	-3.45	0.29	1.82	2.9	1.53	1.8	0.73	0.4	0.28	51.6	48.8	170
28	10	2.8	2.95	-3.22	0.77	1.09	0.3	1.14	0.4	0.88	-0.13	0.26	45	47.3	174
155	56	2.77	2.95	-3.21	0.3	1.37	1.4	1.22	0.8	0.9	0.3	0.2	38.5	38.6	218
66	23	2.87	2.95	-3.16	0.61	0.97	0.1	0.91	0	1	0.12	0.16	42.4	44.7	162
532	182	2.92	2.94	-3.13	0.28	1.54	1.9	1.37	0.8	0.94	0.19	0.2	51.8	51.5	148
24	10	2.4	2.94	-3.09	0.52	0.58	-1.1	0.61	-1	1.53	0.32	0.33	45.7	40.6	69
37	13	2.85	2.94	-3.06	0.75	1.67	1	0.89	0	1.01	0.75	0.2	57.8	47.8	100
99	47	2.11	2.94	-3.02	0.23	0.89	-0.6	0.97	0	1.23	0.57	0.49	40.1	39	118
65	23	2.83	2.93	-2.94	0.54	1.63	1.3	2.1	1.9	0.73	-0.21	0.18	43.8	52.7	114
162	60	2.7	2.93	-2.88	0.28	0.63	-1.9	0.4	0.6	1.4	0.55	0.47	51.8	52.7	120
14	5	2.8	2.93	-2.83	1.07	0.82	0	0.71	0	1.13	0.55	0.13	60	69.8	19
91	37	2.46	2.92	-2.73	0.28	1.04	0.2	1.04	0.2	0.99	0.34	0.25	50	40.8	28
26	10	2.6	2.91	-2.65	0.59	0.97	0	0.84	-0.2	1.23	0.86	0.31	42	41.8	63
94	34	2.76	2.91	-2.6	0.39	1.45	1.3	1.03	0.1	0.98	0.67	0.21	36.9	40	88
62	24	2.58	2.9	-2.57	0.37	1.42	1.4	1.28	0.9	0.75	0.44	0.27	47.2	43.3	200
98	37	2.65	2.9	-2.5	0.32	1.6	2.1	1.5	1.7	0.63	0.22	0.23	40.4	43.8	210
99	34	2.91	2.89	-2.44	0.6	1.54	1	1.1	0.3	0.97	0.34	0.14	11.8	28	13
35	13	2.69	2.89	-2.44	0.57	0.67	-0.6	0.69	-0.6	1.25	0.58	0.25	30.2	39	67
26	10	2.6	2.89	-2.43	0.59	0.57	-0.9	0.6	-0.8	1.42	0.66	0.31	80	52.4	192
88	33	2.67	2.88	-2.37	0.34	1.42	1.4	1.43	1.4	0.68	-0.01	0.24	44.3	43.7	84
8	3	2.67	2.88	-2.36	1.12	0.91	0.1	0.94	0.2	1.01	-0.99	0.06	33.3	33.3	23
26	10	2.6	2.88	-2.33	0.59	1.15	0.4	1	0.1	1.01	0.61	0.31	46.7	45.1	199
68	25	2.72	2.88	-2.31	0.42	1.33	1	1.08	0.3	0.95	0.58	0.2	57.8	47.2	129
59	23	2.57	2.88	-2.29	0.37	1.39	1.3	1.29	1	0.74	0.48	0.24	80	52.4	85
22	9	2.44	2.86	-2.19	0.56	1.42	1	1.57	1.2	0.39	0	0.34	45	41.8	164
74	37	2	2.86	-2.16	0.25	1.04	0.2	1.04	0.2	0.98	0.44	0.26	42.2	39.7	163
20	10	2	2.86	-2.12	0.48	0.34	-2.4	0.34	-2.4	2.28	0.54	0.32	45.7	36.1	74
20	10	2	2.86	-2.12	0.48	0.09	-4.5	0.1	-4.3	2.61	0	0.32	47.1	36.1	80
24	10	2.4	2.84	-2.04	0.52	1.78	1.7	1.83	1.8	-0.1	0.06	0.33	39.8	41.7	176
67	23	2.91	2.84	-2.02	0.73	1.03	0.2	1.22	0.5	0.95	-0.14	0.14	21.7	34	34
25	10	2.5	2.84	-2.01	0.55	1.29	0.7	1.17	0.5	0.71	0.24	0.32	43.3	43.3	95
43	22	1.95	2.84	-2.01	0.32	1.3	1.2	1.3	1.2	0.43	0.19	0.25	40.9	40.3	179
79	31	2.55	2.84	-2.01	0.32	1	0	0.98	0	0.95	0.21	0.34	35.2	43.5	219
325	131	2.48	2.83	-1.97	0.15	1.11	0.9	0.95	-0.2	1.07	0.58	0.42	42.3	43.6	131
60	24	2.5	2.82	-1.86	0.35	1.18	0.7	1.15	0.6	0.8	0.22	0.28	39.2	43.7	143
105	42	2.5	2.8	-1.73	0.27	1.22	1.1	1.11	0.6	0.95	0.61	0.34	33.3	42.2	175
112	43	2.6	2.8	-1.73	0.28	1.27	1.2	1.13	0.6	0.94	0.62	0.26	39.8	45.6	132
76	37	2.05	2.79	-1.7	0.25	0.88	-0.6	0.89	-0.5	1.25	0.33	0.26	43.8	40.2	17
18	10	1.8	2.78	-1.66	0.49	0.52	-1.5	0.55	-1.3	1.67	-0.17	0.3	44.3	32.5	126

53	21	2.52	2.78	-1.65	0.38	0.89	-0.2	0.91	-0.2	0.99	-0.06	0.3	47.1	42.7	76
85	36	2.36	2.78	-1.64	0.27	1.13	0.7	1.12	0.6	0.79	0.19	0.26	44.9	43.5	37
84	34	2.47	2.76	-1.57	0.29	1.18	0.8	1.21	0.9	0.76	0.17	0.26	43.8	51.5	122
62	23	2.7	2.76	-1.57	0.43	0.78	-0.5	0.76	-0.6	1.16	0.38	0.22	30.9	29.5	186
64	23	2.78	2.76	-1.54	0.49	1.14	0.4	0.89	-0.1	1.05	0.6	0.19	69.6	67.8	92
56	28	2	2.75	-1.52	0.35	0.86	-0.4	0.81	-0.5	1.14	0.78	0.76	42.5	42.1	91
294	134	2.19	2.73	-1.42	0.14	0.81	-1.9	0.8	-1.7	1.21	0.43	0.47	39.8	38.1	193
115	41	2.8	2.73	-1.42	0.39	1.06	0.2	0.78	0.9	1.04	0.3	0.29	33.9	35.4	6
56	23	2.43	2.73	-1.39	0.35	1.06	0.3	1.02	0.1	0.91	0.26	0.26	16.7	42.9	79
65	24	2.71	2.73	-1.39	0.43	1.9	2.2	1.93	2.1	0.53	0.18	0.25	16.7	23.2	203
261	122	2.14	2.73	-1.39	0.14	0.93	-0.6	0.94	-0.5	1.02	0.28	0.42	47.9	42.9	212
39	23	1.7	2.73	-1.39	0.33	0.59	-1.9	0.6	-1.8	1.73	0.51	0.23	37.8	39.3	48
81	34	2.38	2.73	-1.39	0.28	1.14	0.7	1.17	0.8	0.81	0.28	0.27	26.5	18.7	81
43	20	2.15	2.72	-1.37	0.34	0.98	0	0.96	0	1.13	0.52	0.31	39.6	37.4	38
94	37	2.54	2.72	-1.37	0.29	1.29	1.2	1.22	0.9	0.86	0.6	0.25	51.4	48.2	110
70	37	1.89	2.72	-1.35	0.25	1.52	2.5	1.59	2.7	-0.15	-0.2	0.26	35.1	33.4	20
50	23	2.17	2.71	-1.33	0.32	1.03	0.1	1.03	0.1	0.98	0.34	0.25	41.9	39.8	188
93	37	2.51	2.69	-1.22	0.29	1.23	1	1.18	0.8	0.82	0.39	0.25	41.9	40.2	216
280	128	2.19	2.68	-1.21	0.14	0.95	-0.4	1.03	0.2	1	0.49	0.53	37.5	41	73
66	36	1.83	2.68	-1.2	0.25	1.45	2.1	1.46	2.2	0.1	-0.02	0.26	34.7	40.5	142
48	18	2.67	2.67	-1.16	0.46	1.28	0.8	1.52	1.3	0.78	0.08	0.23	33.3	35.5	27
23	10	2.3	2.67	-1.15	0.5	0.55	-1.3	0.56	-1.2	1.63	0.14	0.33	28.6	43.7	45
288	124	2.32	2.67	-1.14	0.14	1.6	4.8	1.57	4.4	0.11	0.2	0.34	35	40.3	146
27	13	2.08	2.66	-1.12	0.42	0.86	-0.3	0.86	-0.3	1.25	0.18	0.28	20	35.9	184
68	26	2.62	2.66	-1.12	0.37	1.3	1	1.15	0.5	0.92	0.67	0.22	34.8	34.1	190
68	33	2.06	2.66	-1.11	0.26	0.73	-1.4	0.72	-1.5	1.54	0.35	0.27	46.6	38.4	22
91	37	2.46	2.64	-1.06	0.28	1.01	0.1	1.01	0.1	1.05	0.4	0.25	44.6	40.8	102
62	23	2.7	2.64	-1.03	0.43	1.06	0.2	0.96	0	1.02	0.44	0.22	47.8	47	50
101	37	2.73	2.64	-1.03	0.35	1.41	1.3	1.69	2	0.76	0.03	0.21	55.4	54.2	3
101	37	2.73	2.64	-1.03	0.35	1.04	0.2	0.95	0	1.02	0.34	0.21	62.2	54.2	107
100	37	2.7	2.62	-0.98	0.34	1.45	1.5	1.3	1	0.81	0.32	0.22	44.6	48.5	169
47	24	1.96	2.62	-0.98	0.31	0.27	-4.5	0.28	-4.4	2.46	0.73	0.29	35	38.8	147
20	10	2	2.6	-0.92	0.48	1.23	0.7	1.3	0.8	0.32	-0.38	0.32	30	33.8	82
50	19	2.63	2.6	-0.91	0.44	0.94	0	0.85	-0.3	1.14	0.58	0.26	25	28	1
103	37	2.78	2.6	-0.9	0.39	1.46	1.3	1.4	1.1	0.84	0.2	0.2	35.1	43.9	83
76	37	2.05	2.59	-0.89	0.25	0.78	-1.2	0.79	-1.1	1.44	0.31	0.26	32.1	37.4	40
61	23	2.65	2.58	-0.86	0.4	1.36	1.1	1.42	1.2	0.65	-0.25	0.22	52.2	47.3	49
68	26	2.62	2.57	-0.83	0.37	1.28	1	1.28	0.9	0.8	0.24	0.22	63.3	61.3	116
69	26	2.65	2.57	-0.8	0.38	1.32	1	1.29	0.9	0.84	0.33	0.21	71.4	59.7	12
54	21	2.57	2.55	-0.76	0.4	1.23	0.8	1.15	0.5	0.87	0.41	0.29	66.7	56.7	206
47	21	2.24	2.53	-0.69	0.34	0.42	-2.9	0.44	-2.7	2.06	0.68	0.32	40	42.7	198
180	80	2.25	2.53	-0.68	0.18	1.23	1.6	1.24	1.5	0.67	0.44	0.47	35.4	40.4	18
44	23	1.91	2.5	-0.61	0.32	0.63	-1.7	0.63	-1.7	1.77	0.5	0.25	46.7	39.3	72
29	13	2.23	2.5	-0.61	0.43	0.8	-0.5	0.8	-0.5	1.31	0.15	0.29	39.4	42.2	136
64	33	1.94	2.49	-0.58	0.26	0.5	-3.1	0.51	-3	1.91	0.31	0.3	19.2	35.2	15
72	34	2.12	2.48	-0.56	0.26	0.73	-1.4	0.73	-1.5	1.48	0.26	0.28	51.9	39.5	98
123	49	2.51	2.47	-0.53	0.25	0.84	-0.8	0.82	-0.9	1.2	0.41	0.26	50	46.9	89
37	15	2.47	2.46	-0.49	0.44	1	0.1	0.97	0	1.03	0.31	0.28	42.4	35.5	159
51	22	2.32	2.43	-0.41	0.34	1.09	0.4	1.1	0.4	0.88	0.34	0.26	33.3	35.6	36
67	36	1.86	2.42	-0.39	0.26	1.61	2.7	1.65	2.9	-0.24	0	0.36	20.6	27.3	86
24	10	2.4	2.4	-0.33	0.52	1.14	0.4	1.26	0.7	0.78	0.2	0.33	44.4	39.2	156
79	37	2.14	2.4	-0.32	0.25	0.8	-1.1	0.79	-1.1	1.39	0.34	0.27	32.4	36.3	191
51	20	2.55	2.39	-0.32	0.41	0.88	-0.3	0.89	-0.2	1.02	0.17	0.38	37.1	40.1	117
59	33	1.79	2.39	-0.3	0.27	0.89	-0.5	0.88	-0.5	1.29	0.53	0.26	0	0	151
24	10	2.4	2.38	-0.29	0.52	0.82	-0.3	0.79	-0.4	1.38	0.7	0.33	48.3	42.6	141
46	23	2	2.36	-0.23	0.31	0.24	-4.8	0.24	-4.7	2.42	0.08	0.25	42.4	36.9	44
17	10	1.7	2.36	-0.23	0.5	0.31	-2.5	0.32	-2.3	2.19	0.73	0.29	33.3	34.8	53
140	77	1.82	2.36	-0.23	0.18	1.28	2	1.3	2.1	0.43	0.08	0.32	36.9	34.4	171
51	23	2.22	2.35	-0.2	0.32	1.02	0.1	1.01	0.1	1.04	0.49	0.25	46.7	40.2	99
90	37	2.43	2.3	-0.08	0.27	1.37	1.7	1.33	1.5	0.45	-0.05	0.26	47.3	47.2	133
37	22	1.68	2.3	-0.08	0.34	0.54	-2.1	0.54	-2.1	1.71	0.19	0.23	52.3	37.1	2
47	22	2.14	2.28	-0.02	0.32	0.73	-1.1	0.72	-1.1	1.58	0.5	0.26	50	40.6	189
134	60	2.23	2.27	-0.01	0.2	0.83	-1.1	0.87	-0.8	1.18	0.27	0.4	50.4	41.5	196
50	26	1.92	2.24	0.05	0.3	0.15	-6.3	0.16	-6.2	2.63	0.37	0.23	47.9	31.8	96

36	23	1.57	2.24	0.07	0.34	1.86	2.8	1.82	2.6	-0.12	0.09	0.21	24.2	42.2	157
87	37	2.35	2.18	0.2	0.26	1.12	0.6	1.09	0.5	0.8	0.17	0.26	47.3	45.7	39
47	23	2.04	2.16	0.25	0.31	0.86	-0.5	0.86	-0.5	1.26	0.2	0.25	36.8	35	10
17	13	1.31	2.14	0.28	0.56	0.8	-0.3	0.76	-0.3	1.15	0.31	0.18	0	0	41
161	74	2.18	2.13	0.31	0.19	0.78	-1.7	0.78	-1.6	1.31	0.47	0.47	42.6	39.7	70
66	34	1.94	2.1	0.38	0.26	0.81	-1	0.81	-0.9	1.39	0.38	0.27	42.2	39.7	213
92	37	2.49	2.1	0.38	0.28	0.9	-0.4	0.9	-0.4	1.04	0.07	0.25	18.3	26.9	178
21	10	2.1	2.07	0.44	0.48	0.22	-3.2	0.22	-3.2	2.41	0.32	0.33	56.3	39.4	134
68	34	2	2.07	0.44	0.26	1.35	1.7	1.34	1.6	0.36	0.31	0.27	34.8	37.6	87
16	10	1.6	2.06	0.46	0.52	0.97	0	0.92	0	1.17	0.49	0.27	16.2	22.9	112
46	23	2	2.05	0.5	0.31	0.41	-3.2	0.41	-3.2	2.13	0.18	0.25	30.4	37.2	90
208	74	2.81	2.04	0.51	0.29	1.71	2.5	1.88	2.5	0.71	0.03	0.25	57.8	54.1	52
89	43	2.07	2.04	0.52	0.23	0.41	-4.4	0.41	-4.4	2.06	0.09	0.27	40.2	36.6	115
35	20	1.75	2.03	0.54	0.37	0.93	-0.1	0.92	-0.2	1.27	0.67	0.51	38.2	36.1	185
19	9	2.11	2.02	0.55	0.5	1.14	0.4	1.15	0.5	0.74	0.3	0.23	44.4	37.5	8
73	34	2.15	2.02	0.57	0.26	0.33	-4.7	0.34	-4.6	2.17	0.12	0.28	47.1	36.6	139
26	13	2	2	0.61	0.42	0.88	-0.2	0.88	-0.3	1.09	-0.23	0.28	22.2	35.3	173
48	24	2	2	0.61	0.31	0.69	-1.4	0.69	-1.4	1.59	0.32	0.29	50	36.8	217
71	34	2.09	1.99	0.62	0.26	1.08	0.4	1.08	0.4	0.75	-0.01	0.28	34.8	36.5	5
24	10	2.4	1.98	0.65	0.52	0.85	-0.2	0.83	-0.3	1.05	-0.21	0.33	46.8	41.5	155
20	10	2	1.97	0.67	0.48	0.09	-4.5	0.1	-4.3	2.61	0	0.32	53.8	39.1	51
20	10	2	1.97	0.67	0.48	1.61	1.5	1.67	1.7	-0.35	-0.2	0.32	37.5	39.1	167
75	39	1.92	1.9	0.82	0.24	0.26	-5.9	0.27	-5.8	2.38	0.32	0.26	47.1	35.1	14
21	13	1.62	1.88	0.86	0.45	1.45	1.3	1.42	1.2	0.33	0.08	0.24	55	40.6	194
19	10	1.9	1.87	0.9	0.48	0.16	-3.7	0.19	-3.4	2.51	0.49	0.31	52.5	38.5	43
23	10	2.3	1.86	0.91	0.5	0.48	-1.6	0.5	-1.5	1.79	0.32	0.33	46.8	40.6	124
56	26	2.15	1.84	0.97	0.3	0.34	-4	0.35	-3.9	2.29	0.6	0.24	21.7	34	127
24	13	1.85	1.82	1.01	0.42	1.47	1.4	1.5	1.4	0.07	-0.02	0.27	7.7	25.5	207
80	37	2.16	1.8	1.06	0.25	0.53	-3	0.54	-2.9	1.85	0.28	0.27	39.2	38.8	101
65	34	1.91	1.77	1.11	0.26	0.36	-4.4	0.37	-4.2	2.13	0.1	0.27	41.2	31	103
22	12	1.83	1.74	1.19	0.44	0.45	-2	0.46	-1.9	2.12	0.75	0.28	25	28	195
15	10	1.5	1.73	1.21	0.54	0.77	-0.5	0.79	-0.3	1.19	0.09	0.25	60	39.1	205
50	23	2.17	1.72	1.24	0.32	0.47	-2.7	0.47	-2.7	1.98	0.36	0.25	34.8	34.7	77
27	16	1.69	1.7	1.28	0.4	0.61	-1.4	0.64	-1.3	1.5	-0.01	0.27	51.4	40.6	111
100	34	2.94	1.7	1.3	0.72	0.97	0.1	0.77	0	1.02	0.2	0.12	59.8	62	65
27	16	1.69	1.68	1.32	0.4	1.14	0.5	1.12	0.4	0.8	0.2	0.27	27.3	20.5	25
26	20	1.3	1.67	1.35	0.46	0.95	0	1.18	0.5	0.91	-0.17	0.19	13	15.8	202
704	284	2.48	1.62	1.48	0.11	1.14	1.6	1.19	1.9	0.84	0.44	0.49	31.9	30.3	220
44	34	1.29	1.62	1.48	0.35	1.25	0.9	1.2	0.7	0.93	0.34	0.18	26.5	18.7	71
29	18	1.61	1.59	1.56	0.38	0.68	-1.1	0.69	-1.1	1.37	0.02	0.24	12.5	24.1	225
36	23	1.57	1.59	1.57	0.34	1.16	0.6	1.19	0.7	0.71	0.03	0.21	21.7	25.5	123
20	10	2	1.57	1.62	0.48	0.61	-1.1	0.62	-1.1	1.85	0.68	0.32	45.6	36.1	183
67	23	2.91	1.51	1.78	0.73	1.96	1.3	1.29	0.6	0.92	0.32	0.14	60.9	61.4	106
222	91	2.44	1.5	1.82	0.21	0.93	-0.4	0.97	0	1.07	0.68	0.67	44.5	42.4	11
19	10	1.9	1.49	1.85	0.48	0.16	-3.7	0.19	-3.4	2.51	0.49	0.31	48.1	34.2	42
94	36	2.61	1.46	1.92	0.31	0.98	0	0.93	-0.1	1.03	0.24	0.24	48.6	50.1	197
45	33	1.36	1.44	2	0.33	0.85	-0.5	0.83	-0.5	1.09	0.11	0.18	29.4	22.2	145
16	9	1.78	1.4	2.1	0.52	0.34	-2.2	0.35	-2.1	2.13	0.45	0.31	45.8	32.1	215
12	10	1.2	1.4	2.13	0.76	0.89	0	0.8	0	1.06	0.23	0.16	15.4	7.6	209
11	10	1.1	1.31	2.44	1.04	0.9	0.1	0.67	0	1.07	0.32	0.12	8.3	7.5	158
45	16	2.81	1.23	2.79	0.62	1.72	1.3	1.8	1.3	0.77	0.11	0.2	54.2	45.7	56
19	16	1.19	1.19	3.04	0.62	0.87	0	0.82	-0.1	1.07	0.28	0.17	27	35.5	161
58	21	2.76	1.18	3.06	0.49	0.73	-0.6	0.62	-0.8	1.21	0.58	0.25	54.5	51	46
52	28	1.86	1.17	3.12	0.29	1.24	1.1	1.22	1	0.64	0.45	0.24	33.8	35.5	221
171	108	1.58	1.17	3.17	0.18	0.91	-0.6	0.79	-1.3	1.15	0.67	0.62	28.8	29.6	78
503	292	1.72	1.17	3.18	0.1	1.1	1.3	1.22	2.3	0.81	0.52	0.59	33.5	35	222
70	24	2.92	1.14	3.33	0.73	0.87	0	0.57	-0.4	1.09	0.46	0.15	64	60.8	64
59	21	2.81	1.14	3.36	0.54	1.77	1.5	1.13	0.4	0.92	0.58	0.23	60	59.9	7
74	26	2.85	1.14	3.36	0.53	0.81	-0.2	0.69	-0.5	1.12	0.53	0.16	65.2	66.6	4
8	3	2.67	1.12	3.54	1.12	0.81	0	0.79	0	1.15	0.6	0.06	0	11.7	58
65	26	2.5	1.09	3.88	0.34	0.76	-0.9	0.78	-0.8	1.21	0.05	0.23	48.9	47.9	21
13	5	2.6	1.08	3.92	0.82	0.7	-0.3	0.67	-0.3	1.32	0.53	0.17	33.3	37.9	66
31	11	2.82	1.07	4.05	0.76	0.71	-0.2	0.55	-0.5	1.2	0.65	0.22	0	0	144
27	11	2.45	1.07	4.11	0.51	0.83	-0.3	0.8	-0.4	1.11	-0.03	0.31	57.6	47.3	29

26	10	2.6	1.05	4.38	0.59	0.88	-0.1	0.86	-0.1	1.05	0.11	0.31	60	59.9	166
65	24	2.71	1.04	4.77	0.43	1.22	0.7	1.07	0.3	0.91	0.24	0.25	66	59.6	140
109	44	2.48	1.03	5.01	0.26	0.88	-0.5	0.89	-0.5	1.2	0.49	0.31	57.5	53.1	180
34	16	2.13	1.03	5.01	0.38	1.53	1.7	1.52	1.7	-0.09	-0.04	0.29	52.1	35.7	172
35	29	1.21	1.03	5.14	0.44	1.7	1.7	1.63	1.4	0.75	0.1	0.16	25.3	26.7	224
44	18	2.44	1.01	5.77	0.39	1.28	0.9	1.21	0.7	0.64	0.16	0.25	46.9	43.5	168
6	3	2	1.01	5.86	0.86	1.38	0.7	1.38	0.7	0.28	0.92	0.08	33.3	38.2	152
38	23	1.65	1.01	5.95	0.33	0.4	-3.1	0.41	-3	2.01	0.67	0.22	9.8	14.7	104
23	10	2.3	1.01	6.06	0.5	0.72	-0.7	0.72	-0.7	1.54	0.63	0.33	40	41.4	30
5	3	1.67	1.01	6.62	0.91	0.62	-0.4	0.61	-0.4	1.6	-0.99	0.08	33.3	38.2	16
6	3	2	1.01	6.68	0.86	0	-3.8	0	-3.8	3.03	0	0.08	0	0	105
36	16	2.25	1.01	6.76	0.39	0.47	-2.2	0.48	-2.1	1.84	0.17	0.29	80	30.4	182
11	8	1.38	1	7.01	0.66	0.88	0	0.9	0	1.04	-0.21	0.14	27.8	15.9	33
6	3	2	1	7.34	0.86	1.59	0.9	1.59	0.9	-0.26	-0.92	0.08	33.3	37	94
14	8	1.75	1	7.42	0.54	0.48	-1.5	0.47	-1.5	1.92	-0.05	0.17	21.1	28.9	93
13	8	1.63	1	8.32	0.57	0.54	-1.2	0.55	-1.1	1.72	0.35	0.16	69.2	37.2	211
7	5	1.4	1	8.44	0.82	0.94	0.1	0.96	0.1	0.94	-0.45	0.18	40	30	61
8	5	1.6	1	8.44	0.72	0.45	-1.1	0.47	-1	1.89	0.82	0.2	60	36.6	68
--------	--------	--------	-------	-------	------	-----	-----	------	-----	-----	------	------	------	-------	----
69.7	30.5	2.26	2.23	0	0.43	0.99	-0.2	0.96	-0.2		0.28				Mea
83	36.2	0.44	0.66	2.82	0.2	0.42	1.8	0.41	1.7		0.31				S.D
83.3	36.3	0.44	0.67	2.83	0.2	0.42	1.8	0.41	1.7		0.31				S.D
--------	--------	--------	-------	-------	------	-----	-----	------	-----	-----	------	------	------	-------	----

```
+----------------------------------------------------------------------------------------------------
------------+
|  Total   Total   Obsvd  Fair(M)|       Model | Infit      Outfit    |Estim.| Correlation |
|
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | N
experience        |
|---------------------------------+--------------+---------------------+------+-------------+---------
------------|
|  9620    4135      2.33   2.64 |  -1.05   .03 |  .98  -.8  1.01   .1 | 1.03 |   .60   .59 | nonmaste
|  2951    1299      2.27   2.64 |  -1.03   .05 | 1.01   .4  1.01   .1 | 1.00 |   .55   .55 | 2 master
|  1294     633      2.04   1.41 |   2.09   .07 | 1.14  2.5  1.23  3.3 |  .81 |   .65   .69 | 1 expert
|---------------------------------+--------------+---------------------+------+-------------+---------
|  4621.7  2022.3    2.21   2.23 |    .00   .05 | 1.05   .7  1.08  1.2 |      |   .60       | Mean
|  3598.5  1518.4     .12    .58 |   1.47   .02 |  .07  1.4   .10  1.5 |      |   .04       | S.D.
|  4407.3  1859.7     .15    .71 |   1.81   .02 |  .08  1.7   .13  1.9 |      |   .05       | S.D.
+----------------------------------------------------------------------------------------------------
------------+
Model, Populn: RMSE .05  Adj (True) S.D. 1.47  Separation 28.79  Strata 38.73  Reliability 1.00
Model, Sample: RMSE .05  Adj (True) S.D. 1.81  Separation 35.27  Strata 47.36  Reliability 1.00
Model, Fixed (all same) chi-square:  1795.3  d.f.: 2  significance (probability): .00
Model,  Random (normal) chi-square:  2.0  d.f.: 1  significance (probability): .16
```

Figure G-1. Judge’s Experience Measurement Report (arranged by MN).

Table G-2

Site Measurement Report (Arranged by MN).

Total Score	Total Count	Obs. Avg.	Fairm Aver	Measure	S. E.	MnSq	ZStd	MnSq	ZStd	Discrm	PtMea	PtExp	Nu	site
166	74	2.24	1.02	-5.1	0.2	1.25	1.5	1.27	1.4	0.76	0.61	0.66	60	60
250	111	2.25	1.04	-4.2	0.1	1.11	0.9	1.12	1	0.84	0.39	0.4	43	43
337	162	2.08	1.09	-3.2	0.1	0.8	-2.2	0.79	-2.1	1.32	0.6	0.53	37	37
177	78	2.27	1.12	-2.9	0.2	1.21	1.3	1.15	0.9	0.93	0.68	0.66	53	53
146	79	1.85	1.15	-2.7	0.1	1.02	0.2	1.04	0.3	0.85	0.28	0.45	11	11
254	111	2.29	1.15	-2.7	0.1	1.1	0.9	1.09	0.8	0.84	0.41	0.47	64	64
213	110	1.94	1.16	-2.6	0.1	1.03	0.3	1.04	0.4	0.93	0.25	0.29	29	29
125	60	2.08	1.23	-2.2	0.2	0.91	-0.5	0.92	-0.5	1.14	0.19	0.26	20	20
347	143	2.43	1.24	-2.2	0.1	1.6	4.5	1.53	3.5	0.43	0.4	0.57	36	36
363	172	2.11	1.25	-2.1	0.1	1.11	1.1	1.05	0.4	0.94	0.64	0.64	54	54
241	110	2.19	1.27	-2	0.1	0.8	-1.8	0.86	-1	1.34	0.55	0.46	62	62
258	110	2.35	1.28	-2	0.1	1.1	0.7	1.07	0.4	0.97	0.64	0.64	22	22
133	68	1.96	1.29	-1.9	0.1	0.71	-2.2	0.71	-2.2	1.48	0.44	0.41	52	52
225	103	2.18	1.31	-1.9	0.1	1.03	0.3	1.01	0	1	0.7	0.7	13	13
349	146	2.39	1.31	-1.8	0.1	1.11	1	1.12	1	0.89	0.4	0.41	2	2
460	195	2.36	1.37	-1.6	0.1	1.11	1.1	1.13	1	0.87	0.62	0.65	65	65
205	83	2.47	1.38	-1.6	0.1	1.07	0.5	1.04	0.2	0.96	0.42	0.42	39	39
267	127	2.1	1.39	-1.5	0.1	1.12	1	1.11	0.8	0.93	0.67	0.67	48	48
311	133	2.34	1.41	-1.5	0.1	1.21	1.8	1.32	2.3	0.81	0.53	0.49	44	44
444	183	2.43	1.43	-1.4	0.1	1.18	1.7	1.04	0.3	0.94	0.61	0.61	47	47
434	181	2.4	1.45	-1.3	0.1	0.79	-2.3	0.81	-1.7	1.21	0.5	0.48	19	19
167	74	2.26	1.49	-1.2	0.2	0.28	-6.7	0.42	-3.3	1.71	0.84	0.63	55	55
276	123	2.24	1.52	-1.1	0.1	1.25	2.1	1.23	1.8	0.7	0.5	0.56	57	57
280	111	2.52	1.55	-1.1	0.1	1.12	0.9	1.16	0.8	0.98	0.54	0.52	16	16
254	111	2.29	1.58	-1	0.1	0.98	-0.1	0.98	-0.1	1.06	0.49	0.47	17	17
164	79	2.08	1.65	-0.8	0.1	0.85	-1.2	0.85	-1.2	1.34	0.46	0.31	61	61
205	101	2.03	1.67	-0.8	0.1	0.9	-0.7	0.88	-0.8	1.1	0.7	0.68	30	30
182	83	2.19	1.73	-0.6	0.1	1.04	0.3	1.02	0.2	1	0.51	0.47	67	67
92	46	2	1.74	-0.6	0.2	1	0	1.01	0.1	0.95	0.58	0.61	31	31
161	84	1.92	1.79	-0.5	0.1	0.67	-2.6	0.67	-2.1	1.38	0.73	0.65	9	9
101	46	2.2	1.79	-0.5	0.2	0.98	0	1.14	0.7	0.97	0.64	0.66	40	40
116	53	2.19	1.81	-0.4	0.2	1.48	2.3	1.58	2.5	0.52	0.57	0.66	34	34
275	147	1.87	1.81	-0.4	0.1	0.62	-4.2	0.6	-3.6	1.5	0.74	0.62	7	7
248	109	2.28	1.82	-0.4	0.1	0.66	-3.1	0.7	-2.3	1.42	0.63	0.54	46	46
241	86	2.8	1.85	-0.3	0.2	1.32	1.3	0.95	0	0.97	0.39	0.36	25	25
259	101	2.56	1.87	-0.3	0.1	0.96	-0.2	0.97	-0.1	1.06	0.55	0.51	8	8
194	80	2.42	1.89	-0.2	0.1	1.14	0.9	1.14	0.9	0.82	0.41	0.49	66	66
204	75	2.72	1.93	-0.2	0.2	1.06	0.3	1	0.1	1.01	0.39	0.36	38	38
237	102	2.32	1.93	-0.2	0.1	0.45	-5.2	0.6	-2	1.52	0.76	0.62	58	58
279	111	2.51	1.96	-0.1	0.1	1.34	2.5	1.3	2	0.63	0.17	0.32	35	35
289	111	2.6	1.99	-0	0.1	1.17	1.2	1.24	1.5	0.87	0.34	0.38	26	26
123	49	2.51	2.1	0.22	0.2	0.44	-3.3	0.76	-0.6	1.39	0.74	0.61	1	1
274	111	2.47	2.17	0.38	0.1	0.81	-1.5	0.92	-0.5	1.18	0.49	0.45	59	59
398	186	2.14	2.23	0.53	0.1	0.7	-3.5	0.74	-2.5	1.26	0.68	0.64	28	28
299	111	2.69	2.62	1.59	0.2	1.16	0.9	1.71	1.1	0.81	0.43	0.48	27	27
280	110	2.55	2.82	2.46	0.1	1.22	1.4	1.44	2.1	0.78	0.43	0.52	56	56
21	18	1.17	2.84	2.6	0.6	1.8	1.4	2.36	1.9	0.65	-0.3	0.12	21	21
34	29	1.17	2.85	2.68	0.4	1.29	0.7	1.02	0.1	0.97	0.31	0.15	12	12
77	47	1.64	2.91	3.24	0.2	1.47	2.2	1.55	2.3	0.23	0.15	0.47	32	32
40	29	1.38	2.94	3.64	0.3	1.1	0.4	1.12	0.5	0.87	0.04	0.2	63	63

339	129	2.63	2.96	4.02	0.2	0.73	-1.7	0.66	0.1	1.16	0.67	0.63	50	50
88	44	2	2.96	4.11	0.2	1.32	1.6	1.46	2.3	0.32	0.27	0.5	5	5
29	18	1.61	2.97	4.28	0.3	1.36	1.2	1.67	2	0.27	-0.3	0.22	42	42
261	108	2.42	2.97	4.29	0.1	1.17	1.2	1.14	0.9	0.86	0.54	0.59	51	51
100	46	2.17	2.98	4.62	0.2	1.08	0.5	1.07	0.4	0.87	0.48	0.52	45	45
101	46	2.2	2.98	4.72	0.2	1	0	0.97	-0.1	1	0.5	0.51	41	41
289	141	2.05	2.98	4.85	0.1	1.07	0.6	1.08	0.6	0.99	0.72	0.72	15	15
34	18	1.89	2.98	4.96	0.3	0.78	-0.7	0.79	-0.7	1.35	0.03	0.25	14	14
257	94	2.73	2.98	4.97	0.2	0.88	-0.6	0.9	-0.1	1.11	0.55	0.49	4	4
53	28	1.89	2.98	4.99	0.2	1	0	1	0	0.99	0.2	0.24	18	18
107	46	2.33	2.98	5.04	0.2	1.06	0.3	0.96	-0.1	1.2	0.75	0.52	3	3
90	34	2.65	2.99	5.67	0.3	0.58	-1.6	0.39	1	1.42	0.59	0.52	6	6
304	117	2.6	2.99	5.74	0.1	1.1	0.6	1.09	0.4	0.93	0.58	0.61	24	24
147	58	2.53	2.99	5.93	0.2	1.13	0.7	1.33	1.4	0.67	0.2	0.47	33	33
143	60	2.38	3	6.33	0.2	0.68	-1.8	0.82	-0.6	1.21	0.74	0.69	49	49
48	18	2.67	3	7.01	0.4	0.97	0	1.02	0.1	0.93	-0.2	0.23	23	23
166	74	2.24	1.02	-5.1	0.2	1.25	1.5	1.27	1.4	0.76	0.61	0.66	60	60
250	111	2.25	1.04	-4.2	0.1	1.11	0.9	1.12	1	0.84	0.39	0.4	43	43
337	162	2.08	1.09	-3.2	0.1	0.8	-2.2	0.79	-2.1	1.32	0.6	0.53	37	37
177	78	2.27	1.12	-2.9	0.2	1.21	1.3	1.15	0.9	0.93	0.68	0.66	53	53
146	79	1.85	1.15	-2.7	0.1	1.02	0.2	1.04	0.3	0.85	0.28	0.45	11	11
254	111	2.29	1.15	-2.7	0.1	1.1	0.9	1.09	0.8	0.84	0.41	0.47	64	64
164	79	2.08	1.65	-0.8	0.1	0.85	-1.2	0.85	-1.2	1.34	0.46	0.31	61	61
205	101	2.03	1.67	-0.8	0.1	0.9	-0.7	0.88	-0.8	1.1	0.7	0.68	30	30
182	83	2.19	1.73	-0.6	0.1	1.04	0.3	1.02	0.2	1	0.51	0.47	67	67
92	46	2	1.74	-0.6	0.2	1	0	1.01	0.1	0.95	0.58	0.61	31	31
161	84	1.92	1.79	-0.5	0.1	0.67	-2.6	0.67	-2.1	1.38	0.73	0.65	9	9
101	46	2.2	1.79	-0.5	0.2	0.98	0	1.14	0.7	0.97	0.64	0.66	40	40
116	53	2.19	1.81	-0.4	0.2	1.48	2.3	1.58	2.5	0.52	0.57	0.66	34	34
275	147	1.87	1.81	-0.4	0.1	0.62	-4.2	0.6	-3.6	1.5	0.74	0.62	7	7
248	109	2.28	1.82	-0.4	0.1	0.66	-3.1	0.7	-2.3	1.42	0.63	0.54	46	46
241	86	2.8	1.85	-0.3	0.2	1.32	1.3	0.95	0	0.97	0.39	0.36	25	25
259	101	2.56	1.87	-0.3	0.1	0.96	-0.2	0.97	-0.1	1.06	0.55	0.51	8	8
194	80	2.42	1.89	-0.2	0.1	1.14	0.9	1.14	0.9	0.82	0.41	0.49	66	66
204	75	2.72	1.93	-0.2	0.2	1.06	0.3	1	0.1	1.01	0.39	0.36	38	38
237	102	2.32	1.93	-0.2	0.1	0.45	-5.2	0.6	-2	1.52	0.76	0.62	58	58
279	111	2.51	1.96	-0.1	0.1	1.34	2.5	1.3	2	0.63	0.17	0.32	35	35
289	111	2.6	1.99	-0	0.1	1.17	1.2	1.24	1.5	0.87	0.34	0.38	26	26
123	49	2.51	2.1	0.22	0.2	0.44	-3.3	0.76	-0.6	1.39	0.74	0.61	1	1
274	111	2.47	2.17	0.38	0.1	0.81	-1.5	0.92	-0.5	1.18	0.49	0.45	59	59
398	186	2.14	2.23	0.53	0.1	0.7	-3.5	0.74	-2.5	1.26	0.68	0.64	28	28
299	111	2.69	2.62	1.59	0.2	1.16	0.9	1.71	1.1	0.81	0.43	0.48	27	27
280	110	2.55	2.82	2.46	0.1	1.22	1.4	1.44	2.1	0.78	0.43	0.52	56	56
21	18	1.17	2.84	2.6	0.6	1.8	1.4	2.36	1.9	0.65	-0.3	0.12	21	21
34	29	1.17	2.85	2.68	0.4	1.29	0.7	1.02	0.1	0.97	0.31	0.15	12	12
77	47	1.64	2.91	3.24	0.2	1.47	2.2	1.55	2.3	0.23	0.15	0.47	32	32
40	29	1.38	2.94	3.64	0.3	1.1	0.4	1.12	0.5	0.87	0.04	0.2	63	63
339	129	2.63	2.96	4.02	0.2	0.73	-1.7	0.66	0.1	1.16	0.67	0.63	50	50
88	44	2	2.96	4.11	0.2	1.32	1.6	1.46	2.3	0.32	0.27	0.5	5	5
29	18	1.61	2.97	4.28	0.3	1.36	1.2	1.67	2	0.27	-0.3	0.22	42	42
261	108	2.42	2.97	4.29	0.1	1.17	1.2	1.14	0.9	0.86	0.54	0.59	51	51
100	46	2.17	2.98	4.62	0.2	1.08	0.5	1.07	0.4	0.87	0.48	0.52	45	45
101	46	2.2	2.98	4.72	0.2	1	0	0.97	-0.1	1	0.5	0.51	41	41
289	141	2.05	2.98	4.85	0.1	1.07	0.6	1.08	0.6	0.99	0.72	0.72	15	15
34	18	1.89	2.98	4.96	0.3	0.78	-0.7	0.79	-0.7	1.35	0.03	0.25	14	14
257	94	2.73	2.98	4.97	0.2	0.88	-0.6	0.9	-0.1	1.11	0.55	0.49	4	4
53	28	1.89	2.98	4.99	0.2	1	0	1	0	0.99	0.2	0.24	18	18
107	46	2.33	2.98	5.04	0.2	1.06	0.3	0.96	-0.1	1.2	0.75	0.52	3	3
90	34	2.65	2.99	5.67	0.3	0.58	-1.6	0.39	1	1.42	0.59	0.52	6	6

304	117	2.6	2.99	5.74	0.1	1.1	0.6	1.09	0.4	0.93	0.58	0.61	24	24
147	58	2.53	2.99	5.93	0.2	1.13	0.7	1.33	1.4	0.67	0.2	0.47	33	33
143	60	2.38	3	6.33	0.2	0.68	-1.8	0.82	-0.6	1.21	0.74	0.69	49	49
48	18	2.67	3	7.01	0.4	0.97	0	1.02	0.1	0.93	-0.2	0.23	23	23
210	91.9	2.23	2.01	0.6	0.2	1.02	0	1.05	0.2	0.47	Mean	(Count:	66	
107	44.4	0.34	0.71	2.98	0.0	0.27	1.9	0.31	1.4	0.24	S.D.	(Populati	10	
108	44.7	0.34	0.71	3	0.0	0.27	1.9	0.31	1.4	0.24	S.D.	(Sample)	10	
Popul	RMS	0.23	Adj	(True)	S.	2.97	Separatio	12.8	Stra	17.4	Reliabil	0.99		
Sampl	RMS	0.23	Adj	(True)	S.	2.99	Separatio	12.9	Stra	17.5	Reliabil	0.99		
Fixed	(all	same	chi-	12771	d.f.	65	significan	(probabili						
Rando	(norm	chi-	64.5	d.f.:	64	significa	(probabili	0.46						

```
+-----------------------------------------------------------------------------------------
------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Correlation |
|
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Nu item
|
|--------------------------------+--------------+----------------------+------+-------------+---------
|   294     117      2.51   2.54 |   -.73   .18 | 1.17  1.2  1.29  1.1 |  .75 |   .46   .57 |  3 PI3
|
|   403     167      2.41   2.47 |   -.52   .14 |  .85 -1.4   .85  -.3 | 1.18 |   .63   .58 | 26 G3
|
|   645     266      2.42   2.46 |   -.49   .11 | 1.00   .0  1.13   .5 |  .94 |   .53   .56 | 17 N6
|
|   306     123      2.49   2.46 |   -.49   .17 |  .89  -.8   .75  -.5 | 1.22 |   .59   .52 | 34 G14
|
|   634     264      2.40   2.43 |   -.40   .11 |  .72 -3.6   .69 -1.2 | 1.36 |   .66   .56 | 21 N11
|
|   281     117      2.40   2.40 |   -.34   .17 |  .60 -3.6   .56 -2.4 | 1.49 |   .74   .59 | 11 PI11
|
|   281     117      2.40   2.39 |   -.31   .17 |  .89  -.8   .84  -.7 | 1.19 |   .65   .58 |  1 PI1
|
|   617     261      2.36   2.39 |   -.30   .11 |  .92  -.9   .87  -.4 | 1.12 |   .59   .57 | 18 N7
|
|   386     165      2.34   2.38 |   -.27   .14 |  .53 -5.3   .55 -1.7 | 1.54 |   .74   .60 | 37 G17
|
|   624     265      2.35   2.37 |   -.26   .11 |  .95  -.6   .94  -.1 | 1.04 |   .57   .57 | 14 N3
|
|   617     262      2.35   2.37 |   -.26   .11 |  .69 -4.2   .64 -1.5 | 1.40 |   .67   .57 | 15 N4
|
|   619     263      2.35   2.37 |   -.25   .11 |  .79 -2.7   .76  -.9 | 1.30 |   .65   .57 | 13 N2
|
|   297     123      2.41   2.36 |   -.24   .16 |  .73 -2.4   .73  -.6 | 1.34 |   .64   .54 | 29 G9
|
|   297     123      2.41   2.36 |   -.24   .16 |  .98  -.1  1.16   .5 | 1.02 |   .55   .54 | 32 G12
|
|   294     122      2.41   2.35 |   -.21   .16 | 1.08   .6  1.07   .3 |  .97 |   .55   .54 | 31 G11
|
|   610     264      2.31   2.31 |   -.11   .11 |  .86 -1.8   .91  -.3 | 1.11 |   .58   .57 | 20 N10
|
|   608     265      2.29   2.30 |   -.08   .11 |  .76 -3.2   .70 -1.4 | 1.29 |   .64   .58 | 12 N1
|
|   377     166      2.27   2.30 |   -.07   .14 | 1.32  2.8  1.43  1.5 |  .61 |   .52   .61 | 28 G7
|
|   256     113      2.27   2.29 |   -.06   .16 | 1.70  4.7  1.81  3.7 |  .07 |   .38   .61 | 22 S1
|
|   272     118      2.31   2.28 |   -.03   .16 |  .86 -1.1   .78 -1.2 | 1.24 |   .67   .60 |  8 PI8
|
|   285     122      2.34   2.27 |    .00   .16 | 1.09   .7  1.06   .2 |  .98 |   .57   .56 | 30 G10
|
|   270     118      2.29   2.26 |    .02   .16 |  .94  -.4   .91  -.4 | 1.10 |   .63   .60 |  6 PI6
|
|   249     112      2.22   2.23 |    .08   .16 | 1.33  2.4  1.46  2.3 |  .62 |   .53   .61 | 24 S11
|
|   250     113      2.21   2.22 |    .10   .16 | 1.32  2.4  1.47  2.4 |  .51 |   .47   .61 | 23 S8
|
|   367     166      2.21   2.22 |    .11   .14 |  .95  -.4  1.06   .3 |  .99 |   .60   .61 | 25 G1
|
|   266     118      2.25   2.21 |    .13   .16 | 1.11   .9  1.00   .0 |  .89 |   .58   .61 |  2 PI2
|
|   266     118      2.25   2.21 |    .13   .16 |  .95  -.3   .86  -.7 | 1.10 |   .62   .61 |  4 PI4
|
|   282     123      2.29   2.21 |    .14   .16 | 1.26  2.1  1.19   .6 |  .77 |   .55   .57 | 35 G15
|
|   367     167      2.20   2.20 |    .16   .13 |  .92  -.8   .94  -.1 | 1.10 |   .64   .62 | 27 G6
|
|   262     117      2.24   2.19 |    .17   .16 | 1.02   .2   .95  -.2 | 1.03 |   .62   .61 |  5 PI5
|
|   274     122      2.25   2.15 |    .27   .16 | 1.18  1.5  2.10  2.9 |  .62 |   .49   .58 | 33 G13
|
|   255     118      2.16   2.09 |    .40   .16 | 1.18  1.5  1.35  2.0 |  .79 |   .59   .61 | 10 PI10
|
|   267     123      2.17   2.05 |    .50   .15 | 1.40  3.1  1.61  2.0 |  .38 |   .46   .59 | 36 G16
|
|   537     262      2.05   1.98 |    .64   .10 | 1.12  1.4  1.14   .8 |  .77 |   .52   .60 | 16 N5
|
|   245     118      2.08   1.98 |    .64   .16 | 1.06   .5   .97  -.1 | 1.03 |   .64   .61 |  7 PI7
|
|   502     263      1.91   1.81 |   1.04   .10 | 1.44  4.9  1.47  2.8 |  .38 |   .48   .60 | 19 N8
|
|   203     106      1.92   1.77 |   1.13   .16 | 1.58  4.1  1.66  3.5 |  .15 |   .43   .61 |  9 PI9
|
|--------------------------------+--------------+----------------------+------+-------------+---------
|   374.7   164.0    2.28   2.26 |    .00   .14 | 1.03   .0  1.07   .3 |      |   .58       | Mean
(Count: 37)     |
|   145.4    62.7     .14    .17 |    .40   .02 |  .26  2.4   .36  1.5 |      |   .08       | S.D.
(Population)   |
|   147.4    63.6     .14    .17 |    .41   .02 |  .26  2.5   .36  1.5 |      |   .08       | S.D.
(Sample)       |
+----------------------------------------------------------------------------------------------------
------------+
Model, Populn: RMSE .15  Adj (True) S.D. .38  Separation 2.57  Strata 3.76  Reliability .87
Model, Sample: RMSE .15  Adj (True) S.D. .38  Separation 2.61  Strata 3.81  Reliability .87
Model, Fixed (all same) chi-square:  341.8  d.f.: 36  significance (probability): .00
Model,  Random (normal) chi-square:  32.1  d.f.: 35  significance (probability): .61
-----------------------------------------------------------------------------------------------------
-------------
```

Figure G-2. Item Measurement Report (arranged by MN).

APPENDIX H: Websites Used in the Rasch Analysis

Website	Number of Judges Who Evaluated
FootLocker	12
J.Crew	12
JCPenney	12
Walmart	12
Amway	11
Coach	11
OneKingsLane	10
Priceline	10
Nordstrom	9
RestorationHardware	9
BestBuy	8
Dell	8
EddieBauer	8
Expedia	8
NET-A-PORTER	8
SHOP.COM	8
Apple	7
EsteeLauder	7
Nike	7
Target	7
TigerDirect.com	7
Blair	6
Lenovo	6
ABMC	5
Amazon	5
Gilt	5
HP	5
Orbitz	5
OrientalTradingCompany	5
Abercrombie&Fitch	3
AdvanceAutoParts	3
DOJOJP	3
DOTRITA	3
EdibleArrangements	3
Express	3
FAA	3
FDA	3
GAO	3
HRSA	3

KeurigGreenMountain	3
L.L.Bean	3
NASA	3
QVC	3
REI	3
SaksFifthAvenue	3
SEC	3
SSA	3
STATEDEPT	3
TheChildren’sPlace	3
VA	3
Wayfair	3
Williams-Sonoma	3
Adobe	2
AirMac	2
Grainger	2
HarborFreight	2
HomeDepot	2
Macy’s	2
Newsweek	2
Wired	2
BicycleDoctorUSA	1
Chipotle	1
Craigslist	1
eBay	1
Enterprise	1
Etsy	1
Microcenter	1
ULWorkplace	1